# Density Ratio Permutation Tests with connections to distributional shifts and conditional two-sample testing

Alberto Bordino and Thomas B. Berrett

Department of Statistics, University of Warwick

## Abstract

We introduce novel hypothesis tests to allow for statistical inference for density ratios. More precisely, we introduce the Density Ratio Permutation Test (DRPT) for testing $H_0 : g \propto rf$ based on independent data drawn from distributions with densities $f$ and $g$, where the hypothesised density ratio $r$ is a fixed function. The proposed test employs an efficient Markov Chain Monte Carlo algorithm to draw permutations of the combined dataset according to a distribution determined by $r$, producing exchangeable versions of the whole sample and thereby establishing finite-sample validity. Regarding the test's behaviour under the alternative hypothesis, we begin by demonstrating that if the test statistic is chosen as an Integral Probability Metric (IPM), the DRPT is consistent under mild assumptions on the function class that defines the IPM. We then narrow our focus to the setting where the function class is a Reproducing Kernel Hilbert Space, and introduce a generalisation of the classical Maximum Mean Discrepancy (MMD), which we term Shifted-MMD. For continuous data, assuming that a normalised version of $g - rf$ lies in a Sobolev ball, we establish the minimax optimality of the DRPT based on the Shifted-MMD. For discrete data with finite support, we characterise the complex permutation sampling distribution using a noncentral hypergeometric distribution, significantly reducing computational costs. We further extend our approach to scenarios with an unknown shift factor $r$, estimating it from part of the data using Density Ratio Estimation techniques, and derive Type-I error bounds based on estimation error. Additionally, we demonstrate how the DRPT can be adapted for conditional two-sample testing, establishing it as a versatile tool for assessing modelling assumptions on importance weights, covariate shifts and related scenarios, which frequently arise in contexts such as transfer learning and causal inference. Finally, we validate our theoretical findings through experiments on both simulated and real-world datasets.

## 1  Introduction

In modern statistical applications, it is increasingly common to encounter multiple datasets originating from different sources (Koh et al., 2020). Consequently, significant efforts have been devoted to developing methods that effectively combine these datasets to address various inferential tasks (see Storkey, 2009; Weiss et al., 2016, for surveys in transfer learning). For example, we may wish to draw inferences about a test data population, even when some of our training data come from related but distinct distributions. Such challenges are prevalent in many practical scenarios. For instance, in the context of Large Language Models,

arXiv:2505.24529v1 [stat.ME] 30 May 2025

thoughtfully incorporating human-authored data alongside model-generated content during training helps maintain diversity and enhance overall performance (see Ji et al., 2025, for a survey). As another example, in medical applications, practitioners might be interested in making predictions in a particular experimental setting or using specific equipment, while also utilising data collected under different conditions or with alternative tools (see Guan and Liu, 2022, for a survey). To develop statistical procedures that fully utilise the information from all samples, it is crucial to make specific assumptions about the relationship between their distributions. In this regard, a widely adopted and effective approach is to leverage knowledge of the density ratio – commonly known as the *importance* in the importance sampling literature (Tokdar and Kass, 2010). This ratio can be used to implement importance weighting techniques (e.g. Kahn and Harris, 1951; Horvitz and Thompson, 1952; Owen and Zhou, 2000), which assign varying weights to data points to correct for biases and emphasise relevant observations.

More precisely, focusing on the case of two independent collections of i.i.d. samples, suppose we have access to two distinct datasets, each coming from a different distribution, and the primary goal is to combine the information from these datasets effectively. For instance, suppose we have data $(X_1, \ldots, X_n)$ from a distribution $P_f$ on a domain $\mathcal{X}$ and data $(Y_1, \ldots, Y_m)$ from another distribution $P_g$ on a domain $\mathcal{Y} = \mathcal{X}$. To integrate these datasets in a meaningful way, we need to make assumptions about the relationship between the distributions. A powerful and commonly used approach is to assume that we know (or can estimate) the density ratio $g/f$, where $f$ and $g$ are the densities of $P_f$ and $P_g$, respectively, with respect to a common dominating measure. This density ratio serves as a crucial bridge that allows us to connect the two distributions and take advantage of their combined information efficiently. Applications of methodologies based on the density ratio are widespread. They appear in nonparametric regression (Ma et al., 2023), efficient estimation with additional incomplete data (Berrett, 2024), testing under distributional shifts (Thams et al., 2023), quantile function estimation (Chen and Liu, 2013), reinforcement learning (e.g. Sutton and Barto, 2018), and adjusting for confounding or selection bias in causal inference (Robins et al., 2000). Another significant area where density ratios play a vital role is in the machine learning domain, particularly in tasks like domain adaptation and transfer learning, where distribution shifts between training and testing data are common and can significantly degrade prediction accuracy. For a collection of ten datasets that reflect diverse, real-world distribution shifts, see Koh et al. (2020). In such settings, it is commonly assumed that train and test data originate from different but related distributions, with a prominent example being the *covariate shift* assumption (Tibshirani et al., 2019; Jin and Candès, 2023), which attributes the entire distributional shift to differences in the covariates. To formalise this, consider a regression or classification problem with $(X, Y) \in \mathcal{Z}^{\mathrm{cov}} \times \mathcal{Z}^{\mathrm{pred}} \subseteq \mathbb{R}^d \times \mathbb{R}$, where the training set $Z_{\mathrm{train}} = \{(X_i^{\mathrm{tr}}, Y_i^{\mathrm{tr}})\}_{i=1}^{n_{\mathrm{tr}}}$ are i.i.d. samples from $f(x, y)$, and the test set $Z_{\mathrm{test}} = \{(X_i^{\mathrm{test}}, Y_i^{\mathrm{test}})\}_{i=1}^{n_{\mathrm{test}}}$ are i.i.d. samples from $g(x, y)$. Under the covariate shift assumption, the density ratio depends solely on the covariates, meaning there exists a function $r : \mathcal{Z}^{\mathrm{cov}} \to \mathbb{R}^+$ such that $g(x, y)/f(x, y) = r(x)$. This implies that the conditional distribution of the predictor given the covariates remains invariant between the training and testing data (see Section 4.2, Lee et al. (2024), and Hu and Lei (2020)). When this assumption holds, it becomes possible to enhance predictive performance by reweighting the training data according to the marginal density ratio (see, e.g. Gretton et al., 2009a; Sugiyama and Kawanabe, 2012).

While the covariate shift framework is reasonable in many practical applications, real-world scenarios often involve more complex types of distributional shifts. As such, it becomes essential to have a statistical

procedure that can confidently determine whether the covariate shift assumption—or other similar assumptions involving the density ratio—is valid. In this paper, we address the problem of testing the hypothesis

$$H_0 : g \propto r\, f, \tag{1}$$

based on independent data $X_1, \ldots, X_n \sim P_f$ and $Y_1, \ldots, Y_m \sim P_g$ as introduced above. The (unnormalised) density ratio $r$ is treated as fixed here, though in Section 4.1 we extend the scope of our work to include settings where $r$ is hypothesised to lie in a parametric family and is estimated from data. Notably, if $r$ is constant, the hypothesis (1) reduces to that of the classical *two-sample testing* problem, a cornerstone of nonparametric statistics (e.g. Ingster and Suslina, 2003). It is important to emphasise that our focus is not on testing whether $f = g$ based on an estimate of the density ratio; for this related problem, readers may refer to Hido et al. (2011), Sugiyama et al. (2011), and Kanamori et al. (2010b). Instead, our results can be seen as initiating a line of work in statistical inference for density ratios, important objects for which the existing focus is on estimation guarantees. This literature was comprehensively reviewed by Sugiyama et al. (2010); we discuss this work and recent advances in more detail in Section 4.1.

Our approach to assessing (1) centres on the novel Density Ratio Permutation Test (DRPT), which employs a suitable resampling scheme to calibrate any given test statistic. Permutation methods are widely used in hypothesis testing, due to their guaranteed validity and strong power properties in both classical asymptotic (Hoeffding, 1952; Lehmann and Romano, 2006) and non-asymptotic (Berrett et al., 2021; Kim et al., 2022) senses. In two-sample testing permutations are used to approximate null distributions by randomly reassigning observation labels, though our method introduces important modifications to allow their use in our more general problem. Traditionally, permutation tests draw permutations uniformly at random, which, in classical settings, is sufficient to satisfy the *randomisation hypothesis*—as discussed in Lehmann and Romano (2006)—thereby providing exact control of the Type-I error. However, this conventional approach is not appropriate for our testing problem (1) since the two samples are not exchangeable under the null hypothesis. We address this limitation by introducing a modified permutation sampling approach (Algorithm 1), which produces permutations of $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ according to a distribution dependent on the shift function $r$, restoring exchangeability of the true and resampled datasets under $H_0$ and thus ensuring finite-sample validity (Proposition 3). When $r$ is constant, our method naturally reduces to the classical permutation test.

This non-uniform permutation sampling approach represents a departure from standard methodology, but aligns with recent developments tackling other testing problems, including:

1. The tests of equality of distributions, given biased data, presented in Kang and Nelson (2009), which will be discussed later;

2. The Conditional Permutation Test (CPT) developed in Berrett et al. (2020) for testing conditional independence ($H_0 : X \perp\!\!\!\perp Y | Z$), which draws permutations based on the knowledge of the conditional distribution of $X | Z$;

3. The framework proposed by Ramdas et al. (2022) for testing data exchangeability using non-uniform permutation distributions. This approach can be seen as a dual to sampling non-uniform rearrangements of indices, as it samples permutations uniformly but incorporates weights in the definition of

the $p$-value. While this methodology could theoretically be applied in our setting, its practicality is limited by the need for an extremely large number of permutations.

An alternative approach to addressing the problem in (1) would involve using the knowledge of $r$ to perform importance sampling on $(X_1, \ldots, X_n)$, thereby creating a sample from a distribution proportional to $rf$, which can then be combined with $(Y_1, \ldots, Y_m)$ to apply classical two-sample testing methods. This broad strategy has been employed in works such as Thams et al. (2023) and Chau et al. (2024) for other statistical problems. However, we intentionally avoid this approach as importance sampling typically results in a much smaller effective sample size, leading to less powerful tests in practice. We will empirically validate this claim in Section 5.1. Finally, we also mention the fact that our methodology also parallels recent advances in conformal prediction under distributional shifts, where non-exchangeable data scenarios necessitate modified approaches to maintain valid coverage (Tibshirani et al., 2019; Hu and Lei, 2020; Prinster et al., 2024).

Perhaps the most closely related work is Kang and Nelson (2009), which focuses on testing the null hypothesis $f = g$ based on biased data sampled from $w_1 f$ and $w_2 g$, where $w_1$ and $w_2$ are positive functions, instead of directly from $f$ and $g$. Their method aligns with ours in that the $p$-value is computed by sampling permutations according to a specific distribution dependent on $w_1$ and $w_2$. They propose a sampling algorithm for this distribution and analyse its asymptotic validity and power. In comparison, our sampling algorithm (Algorithm 1) preserves the target distribution at each step, and results in exact validity for finite sample sizes (Proposition 3). Additionally, we provide non-asymptotic, minimax rate-optimal power guarantees, going beyond asymptotic results. For more on two-sample testing under biased sampling schemes, we refer readers to Navarro et al. (2003); Kang and Nelson (2008); Economou and Tzavelas (2013, 2014); wen Chang and Wang (2023).

Finally, as mentioned earlier, there is an intriguing connection between our framework and the conditional two-sample testing problem, recently explored in Lee et al. (2024). Suppose we observe two independent samples $\{(X_i^{(1)}, Y_i^{(1)})\}_{i=1}^{n_1}$ and $\{(X_i^{(2)}, Y_i^{(2)})\}_{i=1}^{n_2}$, drawn from distributions $P_{XY}^{(1)} = P_X^{(1)} P_{Y|X}^{(1)}$ and $P_{XY}^{(2)} = P_X^{(2)} P_{Y|X}^{(2)}$, respectively. The objective of conditional two-sample testing is to test whether

$$P_X^{(1)} \left\{ P_{Y|X}^{(1)}(\cdot|X) = P_{Y|X}^{(2)}(\cdot|X) \right\} = 1$$

holds. Assuming all distributions have corresponding densities, the null hypothesis $f_{Y|X}^{(1)}(y \mid x) = f_{Y|X}^{(2)}(y \mid x)$ is equivalent to $f_{XY}^{(1)}(x,y) = \{f_X^{(1)}(x)/f_X^{(2)}(x)\} f_{XY}^{(2)}(x,y)$, where $f_{XY}^{(1)}$ and $f_{XY}^{(2)}$ are the joint densities. This demonstrates that our setup—while more general—encompasses the conditional two-sample testing problem when the shift function $r$ corresponds to the marginal density ratio. Consequently, our methodology can be applied to conditional two-sample testing, offering a novel permutation-calibrated test. Similar to the density-ratio-based tests proposed by Lee et al. (2024), our method involves estimating the marginal density ratio. However, while their approach requires a double asymptotic calibration—estimating both the marginal density ratio and the null distribution of the test statistic—our method achieves an asymptotic calibration by estimating only the marginal density ratio. For further detail and numerical comparisons, see Sections 4.2 and 5.2.

4

## 1.1 Outline

We now briefly outline our main contributions. In Section 2, we introduce a permutation test for the testing problem (1), based on generating non-uniform permutations of the combined dataset according to a distribution determined by $r$. The test uses an efficient Markov Chain Monte Carlo (MCMC) sampler (Algorithms 1 and 2) to generate these permutations, producing exchangeable versions of the data and ensuring finite-sample validity. This is formalised in Theorem 1 and Proposition 3, our main results on validity. Although the permutation distribution is generally complex, in (7) we explicitly characterise it using Fisher's Non-central Hypergeometric distribution in the case of discrete data with finite support, significantly reducing computational costs. These results are discussed in Section 2.1, with further power analyses provided in Appendix B.

Section 3 is dedicated to a power analysis of the DRPT. We first show that, if the test statistic is an empirical estimate of an IPM characterising the null, the DRPT is consistent under mild assumptions on the function class associated to the IPM (Theorem 6). This builds on the fact that the permuted samples asymptotically satisfy the null hypothesis (Lemma 7), in parallel with classical permutation testing theory. We then focus on the case where the function class is a Reproducing Kernel Hilbert Space (RKHS). In Proposition 8, we introduce the Shifted-MMD, a generalisation of the classical MMD which reduces to a scaled version of it when $r$ is constant. For continuous data, assuming that a normalised version of $g - rf$ lies in a Sobolev ball, we prove that the DRPT based on an estimate of the Shifted-MMD with an appropriate kernel achieves minimax optimality (Theorems 9 and 10). This is the central theoretical contribution of Section 3 which, as well as providing an optimal solution to testing problem (1), also establishes the first minimax optimality results for test procedures based on non-uniform permutations. The analytic techniques we have developed for studying the convergence properties of relevant Markov chains may have broader applications, potentially extending to establish non-asymptotic power guarantees for other methodologies, such as the CPT proposed in Berrett et al. (2020).

Section 4 presents extensions of our method: Section 4.1 addresses the case where the shift factor $r$ is unknown, and quantifies the inflation in Type-I error due to density ratio estimation error (Proposition 11). Section 4.2 adapts the DRPT to the conditional two-sample testing problem. Finally, in Section 5, we validate our methodology through a range of numerical experiments. All proofs are provided in Appendix A.

We conclude the Introduction with some notation that is used throughout the paper. We denote by $\mathbb{R}_+$ the set of positive real numbers and by $\mathbb{R}_{\geq 1}$ the set of real numbers greater than or equal to one. We further set $\mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. The symmetric group of all permutations over the set $[n] := \{1, \ldots, n\}$ is denoted by $\mathcal{S}_n$. For a set $A$, we write $\#A$ to denote its cardinality. The maximum and minimum operators are sometimes denoted by $\vee$ and $\wedge$, respectively. We define the support of a function $f$ defined on a domain $\mathcal{X}$ to be the closure of the set where $f$ does not vanish, that is, $\text{supp}\, f := \overline{\{x \in \mathcal{X} : f(x) \neq 0\}}$. The set of bounded and continuous functions on $\mathcal{X}$ is denoted by $\mathcal{C}_b^0(\mathcal{X})$, and $\delta_x$ denotes the Dirac delta measure centred at $x$. Given $1 \leq p < \infty$ and $d \geq 1$, we define the $L^p$ norm of a function $f$ as $\|f\|_p := \left( \int_{\mathbb{R}^d} |f(x)|^p dx \right)^{1/p}$, and the corresponding $L^p(\mathbb{R}^d)$ space as the set of all measurable functions for which this norm is finite. Furthermore, $\|\cdot\|_\infty$ denotes the essential supremum norm, that is, $\|f\|_\infty := \text{ess sup}_{x \in \mathbb{R}^d} |f(x)|$, and $L^\infty(\mathbb{R}^d)$ refers to the set of functions that are bounded almost everywhere. Finally, we use $C^\infty(\mathcal{X})$ to denote the space of infinitely

differentiable functions on a domain $\mathcal{X}$, and $\bar{z}$ for the complex conjugate of $z \in \mathbb{C}$.

## 2 Permutation methodology and validity

Let $\mathcal{X} = \mathcal{Y}$ and $n, m \geq 1$. Suppose we observe a dataset $Z = (X_1, \ldots, X_n, Y_1, \ldots, Y_m) \subseteq \mathcal{X}^n \times \mathcal{Y}^m$, consisting of $n + m$ independent random variables, where $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_f$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} P_g$. Assume further that $P_f$ and $P_g$ are absolutely continuous with respect to a common dominating measure $\mu$. We propose a refined permutation test designed to detect deviations from the null hypothesis

$$H_0 : g \propto r\, f,$$

where the (unnormalised) density ratio $r : \mathcal{X} \to \mathbb{R}_+$ is assumed to be known. Notice that, by the definition of $\mathbb{R}_+$ given above, we have $r(x) > 0$ for all $x \in \mathcal{X}$. For simplicity, we also assume that the supports of $f$ and $g$ are the same. If this is not the case—that is, if $r(x) > 0$ for all $x \in \mathcal{X}$ but $\operatorname{supp} g \neq \operatorname{supp} f$—then the testing problem becomes easier, as we expect to eventually observe signals in the regions corresponding to the symmetric difference of the supports. Similarly, if $\operatorname{supp} r \neq \mathcal{X}$ but $\operatorname{supp} g \subseteq \operatorname{supp} r$, we can restrict the sample space to $\operatorname{supp} r$ and carry out the analysis as before. If this inclusion does not hold, the testing problem again becomes easier. Here, the methodology is to create copies $Z^{(1)}, \ldots, Z^{(H)}$, with $H \geq 1$, that are exchangeable with $Z = (Z_1, \ldots, Z_{n+m}) := (X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ under $H_0$. With such a choice of the $Z^{(h)}$'s, for any test statistic $T(Z)$, a $p$-value can be given by

$$p = \frac{1 + \sum_{h=1}^{H} \mathbb{1}\{T(Z^{(h)}) \geq T(Z)\}}{1 + H}. \tag{2}$$

This leads to a valid test, as shown in Theorem 1, and it remains to find a method to generate such $Z^{(h)}$'s. Given any vector $\mathbf{x} = (x_1, \ldots, x_{n+m})$ and any permutation $\sigma \in \mathcal{S}_{n+m}$, define $\mathbf{x}_\sigma = (x_{\sigma(1)}, \ldots, x_{\sigma(n+m)})$. Specifically, we set

$$Z^{(h)} = Z_{\sigma^{(h)}} \qquad \text{where} \qquad \mathbb{P}\left\{\sigma^{(h)} = \sigma \mid Z\right\} = \frac{\prod_{i \in \{n+1, \ldots, n+m\}} r(Z_{\sigma(i)})}{\sum_{\tilde{\sigma} \in \mathcal{S}_{n+m}} \prod_{i \in \{n+1, \ldots, n+m\}} r(Z_{\tilde{\sigma}(i)})}. \tag{3}$$

To see why this results in exchangeable data under $H_0$, it is sufficient to argue as in Berrett et al. (2020), and consider an equivalent formulation of the permutation scheme. Let $Z_{()} = (Z_{(1)}, \ldots, Z_{(n+m)})$ be the order statistics of $Z$. When $\mathcal{X} \subseteq \mathbb{R}$, we naturally use the standard ordering on $\mathbb{R}$. In the general case, we can select any total ordering on $\mathcal{X}$; the specific choice does not matter, as its sole purpose is to let us examine the set of $Z$-values without needing to know which value is associated with which data point. Define also $Z_{(p)} = (Z_{(p(1))}, \ldots, Z_{(p(n+m))})$ for each $p \in \mathcal{S}_{n+m}$, and let $P \in \mathcal{S}_{n+m}$ be the permutation given by the ranks of the true observed vector $Z$, so that $Z = Z_{(P)}$. Under the null hypothesis that $g \propto r\, f$, we can show that the distribution of the true ranks $P$, conditional on the order statistics $Z_{()}$, is given by

$$\mathbb{P}\left\{P = p \mid Z_{()}\right\} = \frac{\prod_{i \in \{n+1, \ldots, n+m\}} r(Z_{(p(i))})}{\sum_{\tilde{p} \in \mathcal{S}_{n+m}} \prod_{i \in \{n+1, \ldots, n+m\}} r(Z_{(\tilde{p}(i))})}. \tag{4}$$

Furthermore, examining the definition (3) of the DRPT copies $Z^{(1)}, \ldots, Z^{(H)}$, these can equivalently be defined by

$$Z^{(h)} = Z_{(P^{(h)})} \qquad \text{where} \qquad P^{(h)} \mid Z_{()} \text{ is drawn from (4)}.$$

Comparing with (3), we can thus establish the bijection $P^{(h)} = \sigma^{(h)} \circ P$, which shows the equivalence between (3) and (4). Building upon this, we can prove the following result showing that the DRPT constitutes a valid test of $H_0$.

**Theorem 1.** *Assume that $H_0 : g \propto r\, f$ is true. Suppose that $Z^{(1)}, \ldots, Z^{(H)}$ are drawn i.i.d. from the DRPT sampling scheme given in (4). Then the $H + 1$ random variables*

$$\left( Z, Z^{(1)}, \ldots, Z^{(H)} \right)$$

*are exchangeable. In particular, this implies that for any statistic $T : \mathcal{X}^n \times \mathcal{Y}^m \longrightarrow \mathbb{R}$, the p-value defined in (2) is valid, satisfying $\mathbb{P}\{p \leq \alpha\} \leq \alpha$ for any desired Type-I error rate $\alpha \in (0,1)$ when $H_0$ is true.*

In order to run the DRPT, we need to be able to sample permutations $\sigma^{(1)}, \ldots, \sigma^{(H)}$ from the distribution given in (3) or, equivalently, $P^{(1)}, \ldots, P^{(H)}$ from the distribution given in (4). We now address the challenge of generating these samples efficiently and propose Algorithm 1 to tackle this problem. This procedure is similar to the first algorithm presented in Berrett et al. (2020), where the shift function $r$ plays a role analogous to the conditional density $q(\cdot \mid \cdot)$ of $X \mid Z$.

---

**Algorithm 1** Pairwise sampler for the DRPT

---

1: Initial permutation $P_0$, integer $S \geq 1$. Call $K = \min\{n, m\}$.
2: **for** $t \in [S]$ **do**
3:     Sample a vector of couples $\tau_t = \{(i_1^t, j_1^t), \ldots, (i_K^t, j_K^t)\}$ such that $(i_1^t, \ldots, i_K^t)$ are sampled uniformly and without replacement from $[n]$, and $(j_1^t, \ldots, j_K^t)$ are sampled uniformly and without replacement from $\{n+1, \ldots, n+m\}$, and initialise $P_t$ to be a copy of $P_{t-1}$.
4:     **for** $k \in [K]$ **do**
5:         Draw a Bernoulli random variable $B_{i_k, j_k}^t$ with

$$\mathbb{P}\{B_{i_k, j_k}^t = 1\} = \frac{r\big(Z_{(P_{t-1}(i_k^t))}\big)}{r\big(Z_{(P_{t-1}(i_k^t))}\big) + r\big(Z_{(P_{t-1}(j_k^t))}\big)} := p_{i_k, j_k}^t, \tag{5}$$

        and swap $P_t(i_k^t)$ with $P_t(j_k^t)$ if $B_{i_k, j_k}^t = 1$.
6:     **end for**
7: **end for**
8: **return** $P_S$.

---

Algorithm 1 is easily parallelisable due to the disjoint structure of the pairs in $\tau_t$, and, most importantly, it accurately targets the distribution in (4), guaranteeing that the resulting Markov chain converges to the desired stationary distribution, as established in the following proposition.

**Proposition 2.** *For any initial permutation $P_0$, the distribution (4) of the permutation $P$ conditional on $Z_{()}$ is the unique stationary distribution of the Markov chain defined in Algorithm 1.*

**Remark 1.** *By examining the proof of Proposition 2, we see that other proposal mechanisms can also target the invariant distribution in (4). Specifically, Step 5 in Algorithm 1 switches the indices $i_k^t$ and $j_k^t$ with*

probability $p_{i_k,j_k}^t$ as in (5) but the key property used in the proof is

$$\frac{\mathbb{P}\{B_{i_k,j_k}^t = 1\}}{\mathbb{P}\{B_{i_k,j_k}^t = 0\}} = \frac{r\big(Z_{(P_{t-1}(i_k^t))}\big)}{r\big(Z_{(P_{t-1}(j_k^t))}\big)}.$$

Thus, any alternative distribution on $B_{i_k,j_k}^t$ satisfying this ratio would work equally well. For example, for the theoretical analysis of the methodology, it is more convenient to consider

$$\tilde{p}_{i_k,j_k}^t := \frac{\widehat{\lambda} mn\, r\big(Z_{(P_{t-1}(i_k^t))}\big)}{\big\{n + \widehat{\lambda} mr\big(Z_{(P_{t-1}(i_k^t))}\big)\big\}\big\{n + \widehat{\lambda} mr\big(Z_{P_{t-1}(j_k^t)}\big)\big\}} \tag{6}$$

with $\widehat{\lambda}$ such that $\sum_{i=1}^n \frac{\widehat{\lambda} mr(X_i)}{n + \widehat{\lambda} mr(X_i)} = \sum_{j=1}^m \frac{n}{n + \widehat{\lambda} mr(Y_j)}$.

**Remark 2.** *Testing $H_0 : g \propto rf$ is equivalent to testing $H_0' : f \propto \frac{1}{r} g$. We now show that our method is similarly invariant under taking reciprocals of $r$ and relabelling the samples, using either $p_{i_k,j_k}^t$ or $\tilde{p}_{i_k,j_k}^t$. For $p_{i_k,j_k}^t$, observe that the appropriate quantity after relabelling is*

$$(p_{i_k,j_k}^t)' := \frac{1/r\big(Z_{(P_{t-1}(j_k^t))}\big)}{1/r\big(Z_{(P_{t-1}(i_k^t))}\big) + 1/r\big(Z_{(P_{t-1}(j_k^t))}\big)} = \frac{r\big(Z_{(P_{t-1}(i_k^t))}\big)}{r\big(Z_{(P_{t-1}(i_k^t))}\big) + r\big(Z_{(P_{t-1}(j_k^t))}\big)},$$

*which matches $p_{i_k,j_k}^t$ exactly. In the case of $\tilde{p}_{i_k,j_k}^t$, define*

$$(\tilde{p}_{i_k,j_k}^t)' := \frac{\widehat{\theta} nm/r\big(Z_{(P_{t-1}(j_k^t))}\big)}{\big\{m + \widehat{\theta} n/r\big(Z_{(P_{t-1}(i_k^t))}\big)\big\}\big\{m + \widehat{\theta} n/r\big(Z_{(P_{t-1}(j_k^t))}\big)\big\}},$$

*with $\widehat{\theta}$ such that $\sum_{j=1}^m \frac{\widehat{\theta} n/r(Y_j)}{m + \widehat{\theta} n/r(Y_j)} = \sum_{i=1}^n \frac{m}{m + \widehat{\theta} n/r(X_i)}$. Here, $n$ and $m$ are switched because the roles of the $X$'s and $Y$'s are interchanged. By algebraic manipulation, one finds that $\widehat{\theta} = \widehat{\lambda}^{-1}$. Substituting back into $(\tilde{p}_{i_k,j_k}^t)'$ then shows $(\tilde{p}_{i_k,j_k}^t)' = \tilde{p}_{i_k,j_k}^t$, confirming the invariance of our algorithm under reciprocal transformations of $r$.*

By Proposition 2, Algorithm 1, when executed for sufficiently many steps $S$, generates a copy $Z_{(P_S)}$ which acts as an appropriate control for $Z$ in testing $H_0$. In fact, we can make a much stronger statement. Under the null hypothesis, the original permutation $P$ conditionally on $Z_{()}$ follows distribution (4). Consequently, initialising Algorithm 1 with $P_0 = P$ (equivalently, with the original data vector $Z$) constitutes initialisation from the stationary distribution. This implies that $Z_{(P_S)}$ represents a draw from the exact target distribution for any value of $S$, thereby providing a valid control for $Z$ regardless of the number of steps taken. However, the statistical power for rejecting the null hypothesis diminishes when $S$ is small, as the control copy exhibits excessive similarity to the original data vector $Z$. For practical implementation, we generate $H$ copies, denoted as $Z^{(h)}$ for $h \in \{1, \ldots, H\}$, with the requirement that the original data and all such $H$ permutations maintain mutual exchangeability. This can be achieved through a star-shaped sampling scheme, following the methodology introduced by Besag and Clifford (1989) and subsequently applied in permutation-based approaches by Berrett et al. (2020) and Ramdas et al. (2022).

Algorithm 2, when initialised with $P_0 = P$, provides an exchangeable sampling mechanism, since the

**Algorithm 2** Star-shaped sampler for the DRPT
___

1: Initial permutation $P_0$, integer $S \geq 1$.
2: Let $P_*$ the output of Algorithm 1 after $S$ steps, initialised at $P_0$.
3: **for** $h \in [H]$, independently **do**
4: ___ Let $P^{(h)}$ the output of Algorithm 1 after $S$ steps, initialised at $P_*$.
5: **end for**
6: **return** $(P^{(1)}, \ldots, P^{(H)})$.
___

permutation $P_*$ lies $S$ steps away from each of the permutations $(P, P^{(1)}, \ldots, P^{(H)})$ and the Markov chain associated to Algorithm 1 is reversible, as shown in the proof of Proposition 2. The following result verifies exchangeability and ensures that the results of Theorem 1 remain satisfied when the permuted vectors $Z^{(1)}, \ldots, Z^{(H)}$ are obtained via Algorithm 2.

**Proposition 3.** *Let $Z_{()}$ and $P$ be the order statistics and ranks of $Z$, as defined previously, so that $Z = Z_{(P)}$. Let $(P^{(1)}, \ldots, P^{(H)})$ be the output of Algorithm 2, when initialised at $P_0 = P$, and let $Z^{(h)} = Z_{(P^{(h)})}$ for each $h = 1, \ldots, H$. If $H_0 : g \propto r f$ is true, the $H + 1$ random variables $\left( Z, Z^{(1)}, \ldots, Z^{(H)} \right)$ are exchangeable.*

## 2.1 The DRPT for discrete data

One significant drawback of permutation tests is their computational cost, which in our setting stems from choosing large values for $H$ in (2) and $S$ in Algorithm 2. In this subsection, we introduce an alternative method for discrete data with finite support that is computationally more efficient, since it enables direct sampling from (4) without relying on the MCMC sampler. Let $\mathcal{X} = \mathcal{Y} = \{0, \ldots, J\} =: \mathcal{J}$ with $J \geq 1$, and let $X_i \overset{\text{i.i.d.}}{\sim} f$ and $Y_i \overset{\text{i.i.d.}}{\sim} g$, where

$$\mathbb{P}_f\{X = j\} = f_j \quad \text{and} \quad \mathbb{P}_g\{Y = j\} = g_j,$$

for all $j \in \mathcal{J}$, with $\sum_{j=0}^{J} f_j = \sum_{j=0}^{J} g_j = 1$. Given a sequence of positive numbers $(r_0, \ldots, r_J)$, the testing problem becomes

$$H_0 : g_j \propto r_j f_j \quad \text{for all } j \in \mathcal{J},$$

so that $r(x) = \sum_{j \in \mathcal{J}} r_j \mathbb{1}\{x = j\}$. At the sample level, the DRPT generates a permutation $p$ given $Z_{()}$ as in (4). Conditioned on the data, each permutation preserves the total number of units in each of the $J + 1$ categories, where the count for category $j \in \mathcal{J}$ is denoted by $\text{tot}_j$. What changes is how these values are split between the first $n$ and the last $m$ data points. We can represent this using the following table:

|   | $Z_{(p(1:n))}$ | $Z_{(p(n+1:n+m))}$ | + |
|---|---|---|---|
| 0 | $\text{tot}_0 - N_{Y,0}^p$ | $N_{Y,0}^p$ | $\text{tot}_0$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $J$ | $\text{tot}_J - N_{Y,J}^p$ | $N_{Y,J}^p$ | $\text{tot}_J$ |
| + | $n$ | $m$ | $m + n$ |

where $N_{Y,j}^p = \#\{i \notin [n] : Z_{(p(i))} = j\} = \mathbb{1}\{Z_{(p(n+1))} = j\} + \ldots + \mathbb{1}\{Z_{(p(n+m))} = j\}$ for all $j \in \mathcal{J}$. If we restrict attention to test statistics that are functions of this frequency table, which is natural given that it is a sufficient statistic in our model, the behaviour of the DRTP can be characterised through the distribution

9

of $(N_{Y,0}^p, \ldots, N_{Y,J}^p)|Z_{()}$. In this regard, given the data and for $w = (w_0, \ldots, w_J)$ such that $\sum_{j=0}^J w_j = m$ and $0 \vee (\mathrm{tot}_j - n) \le w_j \le m \wedge \mathrm{tot}_j$ for all $j \in \mathcal{J}$, we have

$$\mathbb{P}\{N_{Y,0}^p = w_0, \ldots, N_{Y,J}^p = w_J \mid Z_{()}\} = \sum_{p \in \mathcal{S}_{n+m}: (N_{Y,0}^p, \ldots, N_{Y,J}^p) = w} \mathbb{P}\{P = p \mid Z_{()}\}$$

$$\propto \sum_{p \in \mathcal{S}_{n+m}: (N_{Y,0}^p, \ldots, N_{Y,J}^p) = w} \prod_{i \notin [n]} r(Z_{(p(i))}) = \sum_{p \in \mathcal{S}_{n+m}: (N_{Y,0}^p, \ldots, N_{Y,J}^p) = w} \prod_{j \in \mathcal{J}} r_j^{N_{Y,j}^p}$$

$$= \left( \prod_{j \in \mathcal{J}} r_j^{w_j} \right) \#\{p \in \mathcal{S}_{n+m} : (N_{Y,0}^p, \ldots, N_{Y,J}^p) = w\} = n!m! \prod_{j \in \mathcal{J}} r_j^{w_j} \binom{\mathrm{tot}_j}{w_j} \propto \prod_{j \in \mathcal{J}} r_j^{w_j} \binom{\mathrm{tot}_j}{w_j}. \quad (7)$$

Notice that in the penultimate step we used the fact that $\#\{p \in \mathcal{S}_{n+m} : (N_{Y,0}^p, \ldots, N_{Y,J}^p) = w\} = n!m! \prod_{j \in \mathcal{J}} \binom{\mathrm{tot}_j}{w_j}$, as there are $\binom{\mathrm{tot}_j}{w_j}$ ways of choosing $w_j$ many $j$'s for the last $m$ data points for all $j \in \mathcal{J}$. Considering all the possible ways in which we can further permute the first $n$ and last $m$ values gives the extra factor of $n!m!$. This shows that $(N_{Y,0}^p, \ldots, N_{Y,J}^p)|Z_{()}$ is distributed according to *Fisher's Multivariate Noncentral Hypergeometric distribution*, which is a generalisation of the hypergeometric distribution where sampling probabilities are adjusted by weight factors (see, e.g. McCullagh and Nelder, 1989, Section 7). Coming back to the testing problem (1), the previous argument shows that in the case of discrete data with finite support we can avoid sampling permutations from (4), as it is sufficient to sample $(N_{Y,0}^p, \ldots, N_{Y,J}^p)|Z_{()}$ from (7). This can be done efficiently using for example the R-function `rMFNCHypergeo` from R-package `BiasedUrn`.

Compared to the MCMC-based approach, aside from reducing computational runtime, the DRPT copies of the table above generated through i.i.d. draws from (7) are conditionally independent, given the row totals, across different $h \in [H]$, while the $Z^{(h)}$'s in the output of Algorithm 2 exhibit non-zero correlation due to their shared initialisation. Nonetheless, this dependence decreases for larger and larger values of $S$. Regarding its theoretical guarantees, finite-sample validity is ensured by Theorem 1. Additional properties and a detailed power analysis of this version of the DRPT—henceforth referred to as the *discrete DRPT*—are presented in Appendix B.

## 3 Power analysis and the Shifted MMD

### 3.1 General theory and IPMs

We now delve into the theoretical analysis of our methodology, and show that the DRPT is consistent under mild assumptions, and minimax rate-optimal under some additional smoothness conditions. This subsection is dedicated to proving the former claim. A sequence of tests is called consistent against a given class of alternatives if, as the sample size tends to infinity, the test will reject the null with probability 1 under every alternative. Such results can be proved for the classical $r \equiv 1$ case using the results of Hoeffding (1952). An intuitive explanation of these results can be found in the fact that the permuted sample asymptotically satisfies the null hypothesis. More precisely, letting $\sigma$ be uniformly distributed over the symmetric group

and assuming that $n/m \to \tau > 0$, we have

$$\frac{1}{n}\sum_{i=1}^{n}\varphi(Z_{\sigma(i)}) \xrightarrow{\mathbb{P}} \int \varphi h_1 d\mu \quad \text{and} \quad \frac{1}{m}\sum_{j=n+1}^{n+m}\varphi(Z_{\sigma(j)}) \xrightarrow{\mathbb{P}} \int \varphi h_1 d\mu, \tag{8}$$

for all $\varphi \in C_b^0(\mathcal{X})$, where $C_b^0(\mathcal{X})$ is the space of bounded continuous functions on $\mathcal{X}$, and $h_1 = \frac{\tau}{1+\tau}f + \frac{1}{1+\tau}g$. This means that the empirical distribution of the first $n$ permuted samples converges in distribution to the mixture of $f$ and $g$, and since this is true also for the last $m$ permuted samples, we can conclude that $Z_\sigma$ asymptotically satisfies the null.

We will now extend this for general density ratios $r$. To this end, we first show that there exists a unique density $h$ which satisfies the null hypothesis while preserving the distribution of the combined data.

**Lemma 4.** *There exists a unique density $h$ such that*

$$\frac{n}{n+m}h + \frac{m}{n+m}\frac{rh}{\int rhd\mu} = \frac{n}{n+m}f + \frac{m}{n+m}g,$$

*and it is of the form*

$$h = \frac{nf + mg}{n + \lambda_0 mr},$$

*for a suitable constant $\lambda_0 > 0$ such that $\int hd\mu = 1$.*

This, together with the fact that $\lambda_0 = (\int rhd\mu)^{-1}$ (proved in Lemma 4), further shows that $H_0$ is equivalent to $f = h$ a.s., which is in turn equivalent to

$$\frac{mg}{n + \lambda_0 mr} = \frac{\lambda_0 mrf}{n + \lambda_0 mr} \text{ a.s.,} \quad \text{with } \lambda_0 > 0 \text{ such that } \int \frac{nf + mg}{n + \lambda_0 mr}d\mu = 1.$$

This motivates the introduction, at the population level, of an IPM of the form

$$T_{\mathcal{F},r}(f,g) := \sup_{\varphi \in \mathcal{F}}\left|\int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r}\varphi\, d\mu\right|,$$

for a suitable function class $\mathcal{F}$. It follows from Lemma 4 that we have $T_{\mathcal{F},r}(f,g) = 0$ under the null, but in order to have the reverse implication for a full characterisation of the null we need additional assumptions on $\mathcal{F}$. Following similar lines to Gretton et al. (2012), we prove the following lemma.

**Lemma 5.** *Let $\mathcal{F} = \{\varphi \in \mathcal{H} : \|\varphi\|_{\mathcal{H}} \le 1\}$, where $\mathcal{H}$ is a dense subset of $C_b^0(\mathcal{X})$ with respect to $\|\cdot\|_\infty$, and $\|\cdot\|_{\mathcal{H}}$ is a norm on $\mathcal{H}$. Then $T_{\mathcal{F},r}$ characterises $H_0$.*

This naturally leads to the consistency of the DRPT based on the empirical version of $T_{\mathcal{F},r}$, as the next result shows. In the following, we will denote with $N(A, \delta, \|\cdot\|_*)$ the $\delta$-covering number (see, e.g. Wainwright, 2019, Chapter 5) of the set $A$ with respect to the norm $\|\cdot\|_*$.

**Theorem 6.** *Fix $\alpha \in (0,1)$ and $H > \lceil 1/\alpha - 1\rceil$. Let $\mathcal{H}$ be a dense subset of $C_b^0(\mathcal{X})$ with respect to $\|\cdot\|_\infty$, and suppose there exists a universal constant $\gamma > 0$ such that $\|\cdot\|_\infty \le \gamma\|\cdot\|_{\mathcal{H}}$. Further, suppose*

$N\left(\{\|\varphi\|_{\mathcal{H}} \leq 1\}, \delta, \|\cdot\|_{\infty}\right)$ *is finite for all* $\delta > 0$ *and define*

$$T(z_1, \ldots, z_{n+m}) = \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} \frac{1}{n} \left| \sum_{i=1}^{n} \frac{\widehat{\lambda} m r(z_i)}{n + \widehat{\lambda} m r(z_i)} \varphi(z_i) - \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r(z_j)} \varphi(z_j) \right|, \tag{9}$$

*where* $\widehat{\lambda}$ *satisfies* $\sum_{i=1}^{n+m} \{n + \widehat{\lambda} m r(z_i)\}^{-1} = 1$. *Then, if there exist* $c, C > 0$ *such that* $c \leq r(x) \leq C$ *for all* $x \in \mathcal{X}$ *and* $n/m \to \tau > 0$, *the DRPT based on* $T$ *is consistent, meaning that whenever* $H_0$ *is not true*

$$\mathbb{P}\{DRPT \text{ based on } T \text{ rejects } H_0\} \to 1 \quad \text{as } n, m \to \infty.$$

It is worth noting that the assumption $n \asymp m$ (i.e. that the sample sizes of the two groups are of the same order) is commonly employed in the literature on permutation tests (e.g. Schrab et al., 2023), and two-sample testing (e.g. Li and Yuan, 2024). Additionally, the condition $\|\cdot\|_{\infty} \leq \gamma \|\cdot\|_{\mathcal{H}}$ is not overly restrictive. For instance, it is satisfied when $\mathcal{H}$ is an RKHS based on a uniformly bounded kernel. The proof of Theorem 6 is based on the following lemma, which may be of independent interest, showing that the empirical distribution of the first $n$ permuted samples converges in distribution to the limiting version of $h d\mu$ defined in Lemma 4. This extends (8) for bounded density ratios $r$.

**Lemma 7.** *Let* $\sigma$ *be sampled according to* (3). *Suppose* $n/m \to \tau > 0$ *and that there exist constants* $c, C > 0$ *such that* $c \leq r(x) \leq C$ *for all* $x \in \mathcal{X}$. *Further, define* $h_{\infty} = \frac{\tau f + g}{\tau + \lambda_{\infty} r}$ *for some* $\lambda_{\infty} > 0$ *such that* $\int h_{\infty} d\mu = 1$. *Also define the empirical measure* $\widehat{H}_{n,m} = \sum_{i=1}^{n+m} \{n + \widehat{\lambda} m r(Z_i)\}^{-1} \delta_{Z_i}$, *where* $\widehat{\lambda} > 0$ *is chosen such that* $\sum_{i=1}^{n+m} \{n + \widehat{\lambda} m r(Z_i)\}^{-1} = 1$. *Then for all* $\varphi \in \mathcal{C}_b^0(\mathcal{X})$ *we have*

(i)

$$\mathbb{E}\left[ \left( \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) - \int \varphi d\widehat{H}_{n,m} \right)^2 \right] \leq \frac{\|\varphi\|_{\infty}^2 (c + C)}{c} \frac{n + m}{n^2};$$

(ii)

$$\frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) \xrightarrow{\mathbb{P}} \int \varphi h_{\infty} d\mu.$$

First, notice that $\sum_{i=1}^{n+m} \{n + \widehat{\lambda} m r(Z_i)\}^{-1} = 1$ is equivalent to saying that

$$\sum_{i=1}^{n} \frac{\widehat{\lambda} m r(Z_{\sigma(i)})}{n + \widehat{\lambda} m r(Z_{\sigma(i)})} = \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r(Z_{\sigma(j)})} \tag{10}$$

for all permutations $\sigma \in \mathcal{S}_{n+m}$. This fact is useful throughout our proofs and allows us to see $\widehat{\lambda}$ as a normalisation factor for $r$ in any division of the empirical distribution $\widehat{H}_{n,m}$ into two samples. Furthermore, it is instructive to elaborate on the proof of part (i), as the underlying strategy is novel and serves as a key component in the proof of other results, such as Theorem 9 below. To this end, let $S_n/n := \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) - \int \varphi d\widehat{H}_{n,m}$ denote the relevant quantity where $\sigma$ is sampled from (3), and define $S_n^t/n$ as its analogous counterpart in which $\sigma$ is replaced by $\sigma_t$, the permutation at time $t$ generated by Algorithm 1. If the procedure is initialised at stationarity, we obtain $\mathbb{E}[(n^{-1}S_n)^2] = \mathbb{E}[(n^{-1}S_n^t)^2]$ for all

$t \in \mathbb{N}$. This is particularly advantageous as we can leverage the Markov chain dynamics to find constants $A$ and $B$ such that $\mathbb{E}[(n^{-1}S_n^{t+1})^2 \mid Z, \sigma_t] - (n^{-1}S_n^t)^2 \leq A - B(n^{-1}S_n^t)^2$, where we now see that this left-hand side has mean zero under stationarity. In proving this bound it is very convenient to use $\tilde{p}_{i_k,j_k}^t$ as defined in (6) to relate this zero-mean difference to $S_n^t/n$. This justifies the claim in Remark 1 that, despite its more complex algebraic form, $\tilde{p}_{i_k,j_k}^t$ is better suited than $p_{i_k,j_k}^t$ in (5) for the theoretical analysis of the DRPT. Finally, while Lemma 7 is presented in terms of the first $n$ data points of the permuted sample, the Law of Large Numbers gives an immediate consequence for the last $m$ data points. Indeed,

$$\frac{1}{m} \sum_{j=n+1}^{n+m} \varphi(Z_{\sigma(j)}) = \frac{1}{m} \left\{ \sum_{i=1}^{n} \varphi(X_i) + \sum_{j=n+1}^{n+m} \varphi(Y_j) - \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) \right\}$$

$$\xrightarrow{\mathbb{P}} \tau \int \varphi f d\mu + \int \varphi g d\mu - \tau \int \varphi h_\infty d\mu = \int \varphi \frac{r h_\infty}{\int r h_\infty d\mu} d\mu,$$

since $\lambda_\infty^{-1} = \int r h_\infty d\mu$ as shown in the proof of Lemma 4.

## 3.2 Construction of Shifted MMD

While Theorem 6 holds under mild assumptions, it is theoretical in nature and its practical utility depends on whether the supremum can be explicitly computed and a closed-form expression for $T_{\mathcal{F},r}$ can be derived. To make this feasible, we focus on the specific case where $\mathcal{H}$ is an RKHS, as examined in Gretton et al. (2012). Under this setting, we establish the following representation for the IPM $T_{\mathcal{F},r}$.

**Proposition 8.** *Let $\mathcal{H}$ be an RKHS with measurable kernel $k(\cdot,\cdot)$ with $\int_{\mathcal{X}} \sqrt{k(x,x)} \frac{\lambda_0 r(x)f(x)+g(x)}{n/m+\lambda_0 r(x)} d\mu(x) < \infty$, where $\lambda_0$ is such that $\int \frac{nf+mg}{n+\lambda_0 mr} d\mu = 1$. Then*

$$T_{\mathcal{F},r}^2(f,g) = \left( \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} \left| \int \frac{\lambda_0 r f - g}{n/m + \lambda_0 r} \varphi \, d\mu \right| \right)^2 = \mathbb{E}_{X,X'\sim f} \left[ \frac{\lambda_0^2 r(X)r(X')k(X,X')}{\{n/m+\lambda_0 r(X)\}\{n/m+\lambda_0 r(X')\}} \right]$$

$$+ \mathbb{E}_{Y,Y'\sim g} \left[ \frac{k(Y,Y')}{\{n/m+\lambda_0 r(Y)\}\{n/m+\lambda_0 r(Y')\}} \right] - 2\mathbb{E}_{\substack{X\sim f \\ Y\sim g}} \left[ \frac{\lambda_0 r(X)k(X,Y)}{\{n/m+\lambda_0 r(X)\}\{n/m+\lambda_0 r(Y)\}} \right].$$

First, note that when $r \equiv 1$, it follows that $\lambda_0 = 1$, leading to

$$T_{\mathcal{F},r\equiv1}^2(f,g) = \frac{m^2}{(n+m)^2} \text{MMD}_k^2(f,g),$$

where $\text{MMD}_k(\cdot,\cdot)$ denotes the Maximum Mean Discrepancy (MMD) based on the kernel $k(\cdot,\cdot)$ (see Gretton et al., 2012, Equation (1) and Lemma 6). Consequently, our IPM $T_{\mathcal{F},r}$, which we will now refer to as the shifted-MMD and denote by $\text{MMD}_{r,k}(f,g)$, generalises the MMD to account for the presence of a shift factor, while reducing to a scaled version of it in the case of a constant $r$. A similar quantity to $\text{MMD}_{r,k}(f,g)$, but without the denominators and $\lambda_0$, was proposed in Lee et al. (2024) for conditional two-sample testing, where its definition was motivated by importance weighting. In contrast, we naturally derive $\text{MMD}_{r,k}(f,g)$ from Lemma 4, and see that the presence of the denominator plays a crucial role, as it is linked to the invariance under relabelling the samples in a similar spirit to Remark 2.

To ensure that $\text{MMD}_{r,k}(f,g)$ characterises the null hypothesis, in line with Lemma 5, restrictions on $\mathcal{H}$

must ensure that $\mathrm{MMD}_{r,k}(f,g) = 0$ if and only if $H_0$ holds. This requirement has been extensively studied in the statistical literature, and RKHS's satisfying this property are referred to as characteristic, with their defining kernels termed universal. For a comprehensive discussion see Sriperumbudur et al. (2010) and references therein. Notably, RKHS's induced by Gaussian and Laplacian kernels on $\mathbb{R}^d$ are characteristic.

Turning now to sample versions of the shifted-MMD, applying the same arguments as in Proposition 8 to empirical measures shows that the test statistic in (9) becomes

$$
V(x_1,\ldots,x_n,y_1,\ldots,y_m) = \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{\widehat{\lambda}^2 r(x_i) r(x_j) k(x_i,x_j)}{\{n/m + \widehat{\lambda} r(x_i)\}\{n/m + \widehat{\lambda} r(x_j)\}}
$$
$$
+ \frac{1}{m^2} \sum_{i,j=1}^{m} \frac{k(y_i,y_j)}{\{n/m + \widehat{\lambda} r(y_i)\}\{n/m + \widehat{\lambda} r(y_j)\}} - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\widehat{\lambda} r(x_i) k(x_i,y_j)}{\{n/m + \widehat{\lambda} r(x_i)\}\{n/m + \widehat{\lambda} r(y_j)\}}, \quad (11)
$$

where $\widehat{\lambda}$ satisfies $\sum_{i=1}^{n} \frac{\widehat{\lambda} m r(X_i)}{n + \widehat{\lambda} m r(X_i)} = \sum_{i=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r(Y_i)}$. Ignoring the dependence on $\widehat{\lambda}$, this estimator is a V-statistic and is asymptotically unbiased for $\mathrm{MMD}_{r,k}^2(f,g)$ since, as we show in Lemma 12 in Appendix A.2, we have $\widehat{\lambda} \xrightarrow{\mathbb{P}} \lambda_0$. Following similar reasoning as in Theorem 6, the DRPT using (11) with a kernel $k(\cdot,\cdot)$ from a characteristic RKHS is consistent, provided that $\|\cdot\|_\infty \leq \gamma \|\cdot\|_{\mathcal{H}}$ and $N(\|\varphi\|_{\mathcal{H}} \leq 1, \delta, \|\cdot\|_\infty)$ is finite for all $\delta > 0$. The former condition is verified for uniformly bounded kernels, as the Cauchy-Schwarz inequality and the representer theorem imply $|\varphi(x)| = |\langle \varphi, k(\cdot,x)\rangle_{\mathcal{H}}| \leq \sqrt{k(x,x)}\|\varphi\|_{\mathcal{H}}$. This is a common assumption in the RKHS literature (see, e.g. Gretton et al., 2012, Definition 30 and subsequent results), and it is satisfied by many standard choices of $k(\cdot,\cdot)$. The latter condition is also mild, and it is satisfied for common kernels such as the exponential kernel (e.g. Yang et al., 2020, Lemma D.2).

### 3.3  Minimax rate optimality

While Theorem 6 shows the consistency of the DRPT under mild assumptions, we can further show its minimax rate optimality when $\mathcal{X} = \mathbb{R}^d$ under the extra assumption that $\frac{\lambda_0 r f - g}{n/m + \lambda_0 r}$ lies in a Sobolev ball. For the theoretical analysis, it is more convenient to consider the test statistic

$$
U(x_1,\ldots,x_n,y_1,\ldots,y_m) = \frac{1}{n^2} \sum_{i \neq j=1}^{n} \frac{\widehat{\lambda}^2 r(x_i) r(x_j) k_\zeta(x_i,x_j)}{\{n/m + \widehat{\lambda} r(x_i)\}\{n/m + \widehat{\lambda} r(x_j)\}}
$$
$$
+ \frac{1}{m^2} \sum_{i \neq j=1}^{m} \frac{k_\zeta(y_i,y_j)}{\{n/m + \widehat{\lambda} r(y_i)\}\{n/m + \widehat{\lambda} r(y_j)\}} - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\widehat{\lambda} r(x_i) k_\zeta(x_i,y_j)}{\{n/m + \widehat{\lambda} r(x_i)\}\{n/m + \widehat{\lambda} r(y_j)\}}, \quad (12)
$$

where $k_\zeta(\cdot,\cdot)$ is a multivariate kernel with bandwidth $\zeta \geq 1$ which will be defined later. Note that apart from a slightly unusual normalisation, namely $\frac{1}{n^2}$ and $\frac{1}{m^2}$ in place of $\frac{1}{n(n-1)}$ and $\frac{1}{m(m-1)}$, and the dependence on $\widehat{\lambda}$, (12) is essentially a U-statistic which serves as an estimator of $\mathrm{MMD}_{r,k_\zeta}^2(f,g)$. Coming back to the definition of $k_\zeta$, we consider a characteristic kernel $K : \mathbb{R} \to \mathbb{R}$ belonging to $L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap L^4(\mathbb{R})$ and satisfying $K(0) = 1$ and, given a bandwidth $\zeta \geq 1$, we define a characteristic kernel on $\mathbb{R}^d$ by $k_\zeta(x,y) := \zeta^d \prod_{i=1}^{d} K(\zeta(x_i - y_i))$ for $x,y \in \mathbb{R}^d$. For example, choosing $K_{\mathrm{Gauss}}(u) = e^{-u^2}$ recovers the Gaussian kernel in $\mathbb{R}^d$, while choosing $K_{\mathrm{Lap}}(u) = e^{-|u|}$ yields the Laplace kernel. Further useful properties of $k_\zeta$ can be found in the proof of Theorem 9. For notational convenience, we also define $\varphi_\zeta(u) := \zeta^d \prod_{i=1}^{d} K(\zeta u_i)$ for

$u \in \mathbb{R}^d$, so that $k_\zeta(x, y) = \varphi_\zeta(x - y)$ for all $x, y \in \mathbb{R}^d$. Extensions to the case of separate kernels $K_i(\cdot)$ with distinct bandwidths $\zeta_i$ along each dimension are straightforward and can be obtained by following similar arguments as in Schrab et al. (2023). Moreover, the assumption that $K(0) = 1$ is not essential, as scaling it by a positive constant does not affect the qualitative behaviour of our results; we adopt this assumption to simplify both the computations and the notation.

We now characterise the optimality of the testing procedure within the minimax framework, aiming to identify the smallest separation between the null and alternative hypotheses that allows for reliable discrimination with controlled error. Smoothness is assumed under the alternative but not under the null, as we already showed in Section 2 that the DRPT test guarantees uniform, non-asymptotic control of the Type-I error without requiring any assumptions on the null distribution. For fixed $r : \mathcal{X} \to \mathbb{R}_+$ such that $0 < c \le r(x) \le C$ for all $x \in \mathcal{X}$, and for $\rho > 0$, we are interested in testing

$$H_0 : g \propto rf \quad \text{vs.} \quad H_1^r(\rho) : \sqrt{\frac{n}{m}} \|\psi_r\|_2 > \rho, \tag{13}$$

where

$$\psi_r = \frac{\lambda_0 mrf - mg}{n + \lambda_0 mr} \quad \text{and} \quad \lambda_0 > 0 \text{ is such that} \quad \int \frac{nf + mg}{n + \lambda_0 mr} d\mu = 1,$$

and aim to find the smallest value of $\rho$ such that there exists a test with uniform error control. The choice of this separation is motivated by the fact that $\sqrt{n/m}\|\psi_r\|_2$ is invariant under relabelling the samples, by analogy with Remark 2. We therefore believe this is the most natural form of separation in our setting. Nonetheless, when $r$ is uniformly bounded, alternative choices involving the $L^2$ norm – such as $\|g - \bar{r}f\|_2$ with $\bar{r} = r(\int rf d\mu)^{-1}$ – are also valid, and should be equivalent in terms of the minimax separation rate.

Let $\Psi$ denote the collection of randomised tests based on the combined dataset $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$. Each test is represented by a function $\varphi : \mathcal{X}^n \times \mathcal{Y}^m \to [0, 1]$, where, upon observing $(x_1, \ldots, x_n, y_1, \ldots, y_m)$, the null hypothesis $H_0$ is rejected with probability $\varphi(x_1, \ldots, x_n, y_1, \ldots, y_m)$. Additionally, define $\Psi(\alpha)$ as the subset of $\Psi$ consisting of tests with size $\alpha$, where $\alpha \in (0, 1)$. Define the $d$-dimensional Sobolev ball with smoothness parameter $s > 0$ and radius $L > 0$ as

$$\mathcal{S}_d^s(L) := \left\{ p \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \int_{\mathbb{R}^d} \|\xi\|_2^{2s} |\widehat{p}(\xi)|^2 \, d\xi \le (2\pi)^d L^2 \right\}, \tag{14}$$

where $\widehat{p}$ denotes the Fourier transform of $p$, that is, $\widehat{p}(\xi) := \int_{\mathbb{R}^d} p(x) e^{-i\langle x, \xi \rangle} dx$ for all $\xi \in \mathbb{R}^d$. Throughout this section we will also assume $m \le n \le \tau m$ for some $\tau \in [1, \infty)$, which is to say that $n$ and $m$ are of the same order. For fixed $r : \mathcal{X} \to \mathbb{R}_+$ such that $0 < c \le r(x) \le C$ for all $x \in \mathcal{X}$, we may then define the minimax separation to be

$$\rho_r^* \equiv \rho_r^*(n, m, \theta, \alpha, \beta) := \inf \left\{ \rho > 0 : \alpha + \inf_{\varphi \in \Psi(\alpha)} \sup_{(f,g) \in \mathcal{S}_\theta^r(\rho)} \mathbb{E}_P(1 - \varphi) \le \alpha + \beta \right\},$$

where $P = P_f^{\otimes n} \otimes P_g^{\otimes m}$, $\theta = (d, \tau, M, s, L) \in \mathbb{N}_+ \times \mathbb{R}_{\ge 1} \times \mathbb{R}_+^3$ and

$$\mathcal{S}_\theta^r(\rho) := \left\{ (f, g) : \max(\|f\|_\infty, \|g\|_\infty) \le M, \sqrt{\frac{n}{m}}\|\psi_r\|_2 > \rho \text{ and } \psi_r \in \mathcal{S}_d^s(L) \right\}. \tag{15}$$

We start by stating an upper bound on the minimax separation which is based on the fact that the DRPT using the test statistic (12) with bandwidth $\zeta = n^{\frac{2}{4s+d}}$ has good power when $\rho$ is sufficiently large.

**Theorem 9.** *Let $\mathcal{X} = \mathbb{R}^d$ and fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$. There exists a constant $C_r = C_r(c, C, \theta, \alpha, \beta)$ such that*

$$\rho_r^* \leq C_r n^{-2s/(4s+d)}.$$

Regarding optimality, a matching lower bound has been established for the classical two-sample testing problem over Sobolev balls (e.g. Li and Yuan, 2024, Theorems 3 and 5), showing that there exists a constant $c_1 = c_1(\theta, \alpha, \beta)$ such that

$$\rho_{r \equiv 1}^* \geq c_1 \, n^{-2s/(4s+d)}.$$

The minimax separation rate $\rho_r^*$ will typically depend on the specific choice of $r$. Nevertheless, we derive the following matching lower bound under the assumptions that $r$ is bounded above and below.

**Theorem 10.** *Fix $\alpha, \beta \in (0,1)$ such that $\alpha + \beta < 1$ and suppose that $s \in \mathbb{N}_+$. For all fixed $r : \mathcal{X} \to \mathbb{R}_+$ satisfying $0 < c \leq r(\cdot) \leq C$, there exists a constant $c_r = c_r(c, C, \theta, \alpha, \beta)$ such that*

$$\rho_r^* \geq c_r n^{-2s/(4s+d)}.$$

When $r$ is uniformly bounded, the combined results of Theorems 9 and 10 demonstrate that $\rho_r^* \asymp n^{-2s/(4s+d)}$, confirming that the DRPT with test statistic (12) achieves optimality. Future work could study the dependence of $\rho_r^*$ on $r$, though by analogy with the goodness-of-fit testing problem for densities, this is likely to be technically demanding. Indeed, it is known (e.g. Balakrishnan and Wasserman, 2019) that minimax rates of convergence for testing the null hypothesis that $f = f_0$ given $X_1, \ldots, X_n \sim f$ intricately depend on the specific choice of $f_0$, with the hardest version of the problem being when $f_0$ is a uniform density. Although the simulations in Section 5.1 and the results for binary data in Appendix B seem to suggest that the hardest shifts to test are those closer to $r \equiv 1$, it is still unclear if this is the case in full generality.

Furthermore, the proof of Theorem 10 relies on relating the testing problem (1) to the task of testing whether $f = g$ under a biased sampling scheme in which, rather than observing samples directly from $f$ and $g$, one observes samples from densities proportional to $w_1 f$ and $w_2 g$, where $w_1$ and $w_2$ are positive functions satisfying $w_2/w_1 = r$. This setting, known as two-sample testing under biased sampling schemes, has been studied in the statistical literature (see Kang and Nelson, 2009, and the references therein), and asymptotically valid and powerful tests have been proposed. From a theoretical standpoint, we believe this alternative formulation is equivalent to our original problem in terms of minimax separation, and that similar techniques to those presented in this section may be used to establish non-asymptotic results in that setting.

Finally, we note that the assumption $s \in \mathbb{N}_+$ is common in the statistical literature (Li and Yuan, 2024), and plays a critical role in the lower bound construction. In particular, it ensures that the bump functions introduced in Lemma 14 in Appendix A.2 are orthogonal with respect to the Sobolev inner product $\langle \phi_1, \phi_2 \rangle_{\mathcal{S}_d^s} := \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \, \widehat{\phi_1}(\xi) \, \overline{\widehat{\phi_2}(\xi)} \, d\xi$. This orthogonality holds when $s \in \mathbb{N}_+$ because the bump functions are supported on disjoint sets. However, this argument breaks down for general $s > 0$, as disjoint support no longer guarantees orthogonality in the Sobolev inner product. Extending the construction to non-integer

smoothness may still be possible using techniques similar to those developed in Butucea (2007) and Albert et al. (2022).

## 4 Extensions

### 4.1 Density ratio permutation test with unknown shift function

We move beyond the case of a known shift factor, and consider the harder problem of testing

$$H_0^{\mathcal{R}} : \text{there exists } r_\star \in \mathcal{R} \text{ such that } g \propto r_\star f,$$

where $\mathcal{R}$ is a suitable function class. This can serve as an important preliminary tool for model selection. For instance, it is common to assume that the density ratio belongs to a specific parametric family of functions (Qin, 1998; Kanamori et al., 2010a), with maximum likelihood estimation used for inference. See Qin (1998) and the references therein for a comprehensive overview of applications that rely on this modelling assumption.

The idea here is to split the sample in two disjoint subsets $\mathcal{Z}^{\mathrm{estim}}$ and $\mathcal{Z}^{\mathrm{test}}$: we use $\mathcal{Z}^{\mathrm{estim}}$ to find a good estimate $\widehat{r}$ of $r_\star$, and $\mathcal{Z}^{\mathrm{test}}$ to perform the DRPT defined in Section 2 using $\widehat{r}$ as shift function. In order to find the estimate $\widehat{r}$, we will leverage existing results in the field of Density Ratio Estimation (DRE). A naive approach is to estimate the two densities separately and take the ratio as the proposed estimator, but more direct approaches are known to perform better in practice. Indeed, significant efforts have been made in the recent literature to develop such direct estimators. Various methodologies have been proposed, including the moment matching approach (Gretton et al., 2009b), the probabilistic classification approach (Qin, 1998; Cheng and Chu, 2004; Bickel et al., 2007), and the ratio matching approach (Tsuboi et al., 2009; Kanamori et al., 2010a, 2009, 2012; Sugiyama et al., 2008; Yamada and Sugiyama, 2010). Additionally, other methods have utilised M-estimators combined with non-asymptotic variational characterisations of Csiszár-divergences (Nguyen et al., 2008; Sugiyama et al., 2012). Recent advances have focused on improving robustness (Liu et al., 2017; Choi et al., 2021, 2022) and accommodating missing values (Givens et al., 2023). For a more comprehensive review of DRE methods, we refer readers to Kanamori et al. (2012) and Sugiyama et al. (2012), along with the references therein.

We first prove that if $H_0^{\mathcal{R}}$ is satisfied and $\widehat{r}$ is a good approximation of the true $r_\star$, then the excess Type-I error of the DRPT is bounded by the total variation distance between the product measures of the normalised versions of $\widehat{r} f$ and $r_\star f$. Recall that for any two distributions $P_1, P_2$ defined on the same probability space, the total variation distance is defined as $\mathrm{TV}(P_1, P_2) = \sup_A |P_1(A) - P_2(A)|$, where the supremum is taken over all measurable sets. In the result below, we assume that both the approximation $\widehat{r}$ of the true $r_\star$ and the test statistic $T : \mathcal{X}^n \times \mathcal{Y}^m \to \mathbb{R}$ are deterministic, meaning that they are selected independently of $Z$.

**Proposition 11.** *Assume $H_0^{\mathcal{R}}$ is true, and let $r_\star$ be such that $g \propto r_\star f$. For fixed $H \geq 1$, let $Z^{(1)}, \ldots, Z^{(H)}$ be copies of $Z$ generated from the DRPT (4) or from the exchangeable sampler (Algorithm 2) with fixed parameter $S \geq 1$ using an approximation $\widehat{r}$ of the true $r_\star$. Then, for all $\alpha \in (0,1)$ we have*

$$\mathbb{P}\{p \leq \alpha\} \leq \alpha + \mathrm{TV}\left(\{\bar{r} f\}^{\otimes m}, \{\bar{r}_\star f\}^{\otimes m}\right),$$

where $\bar{r} = \left(\int \widehat{r} f d\mu\right)^{-1} \widehat{r}$, $\bar{r}_\star = \left(\int r_\star f d\mu\right)^{-1} r_\star$ and $p$ is the p-value computed as in (2).

This result ensures that, if $\widehat{r}$ is a good approximation to $r_\star$, then the DRPT will experience at most a mild increase in its Type-I error. Arguing as for the CPT in Berrett et al. (2020, Theorem 4), Proposition 11 is a worst-case result, established with respect to an arbitrary test statistic $T$, which may be chosen adversarially to be maximally sensitive to errors in estimating the true $r_\star$. In practice, however, sensible choices of $T$—such as those in Section 3—may be more robust than the theoretical result implies. Finally, observe that although the total variation is asymmetric due to the presence of $m$ only, this symmetry can be restored by taking the reciprocal of $r_\star$ and arguing in a manner similar to Remark 2. However, since we have assumed $n \asymp m$ throughout the paper, we avoid this additional technicality, as it does not affect the qualitative behaviour of the result.

Regarding the construction of an approximation $\widehat{r}$ of $r_\star$, although in certain settings (such as when the data satisfy the covariate shift assumption) we may have access to a large sample of unlabelled data sufficient to estimate the unknown density ratio, in general we lack such resources and may have to rely on the sample-splitting procedure introduced above. More precisely, assuming $n = m$ for simplicity of presentation, and letting $\mathcal{Z} = (X_1, \ldots, X_n, Y_1, \ldots, Y_n)$ denote the combined sample, we split $\mathcal{Z}$ into two disjoint subsets

$$\mathcal{Z}^{\text{estim}} = (X_1, \ldots, X_N, Y_1, \ldots, Y_N) \quad \text{and} \quad \mathcal{Z}^{\text{test}} = (X_{N+1}, \ldots, X_n, Y_{N+1}, \ldots, Y_n),$$

with $1 \leq N < n$, and use $\mathcal{Z}^{\text{estim}}$ to compute an estimator $\widehat{r}$ of $r_\star$, and $\mathcal{Z}^{\text{test}}$ to perform the DRPT based on $\widehat{r}$. Now, for Proposition 11 to have practical implications, we need to verify that there are settings where we can choose $N$ appropriately so that $r_\star$ can be estimated with high accuracy and the excess Type-I error is guaranteed to be small. More formally, we require that TV $\left(\{\bar{r} f\}^{\otimes(n-N)}, \{\bar{r}_\star f\}^{\otimes(n-N)}\right) = o_{\mathbb{P}}(1)$. To establish this in certain settings of interest, it is possible to leverage results from the DRE literature, which offers a wide range of tools for quantifying how close $\bar{r} f$ is to $\bar{r}_\star f$ under various pseudo-metrics. As an illustrative example, we will make use of Theorem 2 from Nguyen et al. (2008). However, we note that other approaches, such as Sugiyama et al. (2008, Theorem 1 and Example 1) and Kanamori et al. (2012, Theorem 2), may yield similar results under comparable conditions.

**Example 1.** *Suppose $\mathcal{R} = \mathcal{S}_d^s(L)$, the Sobolev ball introduced in (14), with integer $s$ satisfying $s > d/2$. Let $\widehat{r}$ represent the M-estimator, referred to as estimator E1, for $r_\star \in \mathcal{R}$ as introduced in Nguyen et al. (2008). Now, if $0 < c \leq r(\cdot) \leq C$ for all $r \in \mathcal{R}$, Theorem 2 in Nguyen et al. (2008) and the remarks thereafter ensure that $\int (\sqrt{r_\star f} - \sqrt{\widehat{r} f})^2 d\mu = \mathcal{O}_{\mathbb{P}}(N^{-\frac{2s}{2s+d}})$. As a result, writing $\mathrm{H}^2(p, q) = 2(1 - \int \sqrt{pq} d\mu)$ for the squared Hellinger distance between densities $p$ and $q$, we have*

$$\mathrm{H}^2\left(\{\bar{r} f\}, \{\bar{r}_\star f\}\right) = \int (\sqrt{\bar{r}_\star f} - \sqrt{\bar{r} f})^2 d\mu \leq 2\int (\sqrt{\bar{r}_\star f} - \sqrt{\widehat{r} f})^2 d\mu + 2\int (\sqrt{\widehat{r} f} - \sqrt{\bar{r} f})^2 d\mu$$

$$= 2\int (\sqrt{\bar{r}_\star f} - \sqrt{\widehat{r} f})^2 d\mu + 2\left(\sqrt{\int \widehat{r} f d\mu} - 1\right)^2 d\mu \leq 4\int (\sqrt{\bar{r}_\star f} - \sqrt{\widehat{r} f})^2 d\mu = \mathcal{O}_{\mathbb{P}}(N^{-\frac{2s}{2s+d}}),$$

*where in the last inequality we used the Cauchy-Schwarz inequality. This, together with the fact that*

$\text{TV}(p,q) \leq \sqrt{\text{H}^2(p,q)}$ *and* $\text{H}^2(p^{\otimes n}, q^{\otimes n}) = 2\{1 - (1 - \text{H}^2(p,q)/2)^n\}$, *implies that*

$$\text{TV}\left(\{\bar{r}\,f\}^{\otimes(n-N)}, \{\bar{r}_\star f\}^{\otimes(n-N)}\right) \leq \sqrt{\text{H}^2\left(\{\bar{r}\,f\}^{\otimes(n-N)}, \{\bar{r}_\star f\}^{\otimes(n-N)}\right)}$$

$$= \sqrt{2\left\{1 - (1 - \text{H}^2\left(\{\bar{r}\,f\}, \{\bar{r}_\star f\}\right)/2)^{(n-N)}\right\}}$$

$$\leq \sqrt{(n-N)\,\text{H}^2\left(\{\bar{r}\,f\}, \{\bar{r}_\star f\}\right)} + o_\mathbb{P}(1) = O_\mathbb{P}\left(\sqrt{(n-N)N^{-\frac{2s}{2s+d}}}\right),$$

*where in the last inequality we used that fact that* $(1-x)^n \geq 1 - nx$ *for* $n \in \mathbb{N}$ *and* $x \leq 1$. *As a result, we have* $\text{TV}\left(\{\bar{r}\,f\}^{\otimes(n-N)}, \{\bar{r}_\star f\}^{\otimes(n-N)}\right) \lesssim \sqrt{(n-N)N^{-\frac{2s}{2s+d}}}$ *w.h.p., and expect it to vanish as long as* $N$ *is taken sufficiently large, i.e.* $n \ll N^{\frac{2s}{2s+d}}$.

Related results in the literature can be leveraged to derive analogous guarantees for alternative estimators of $r_\star$ across different settings. Broadly speaking, theoretical insights emphasise the importance of choosing the estimation sample size $N$ to be significantly larger than the testing sample size $n - N$ in order to keep the excess Type-I error low. Nonetheless, empirical findings from Section 5 suggest that less extreme splits may perform just as well in practice.

## 4.2 Conditional two-sample testing

As discussed in the introduction, the DRPT can be applied to the conditional two-sample testing problem, which we now revisit. Consider two independent samples: $(X_i^{(1)}, Y_i^{(1)})_{i=1}^{n_1}$ drawn from a joint distribution $P_{XY}^{(1)}$, and $(X_i^{(2)}, Y_i^{(2)})_{i=1}^{n_2}$ drawn from another joint distribution $P_{XY}^{(2)}$, both supported on a product space $\mathcal{X} \times \mathcal{Y}$. Here, $P_{Y|X}^{(1)}$ and $P_{Y|X}^{(2)}$ represent the conditional distributions of $Y^{(1)}$ given $X^{(1)}$ and $Y^{(2)}$ given $X^{(2)}$, while $P_X^{(1)}$ and $P_X^{(2)}$ are the corresponding marginals of $X^{(1)}$ and $X^{(2)}$. The objective is to test whether the conditional distributions are equal, i.e. $P_X^{(1)}\{P_{Y|X}^{(1)}(\cdot|X) = P_{Y|X}^{(2)}(\cdot|X)\} = 1$. Notable contributions to this area include works such as Hu and Lei (2020), Chatterjee et al. (2024), Chen and Lei (2025), Yan et al. (2024), and Huang et al. (2024). Recently, Lee et al. (2024) introduced two general methodologies to address this problem. The first approach converts any conditional independence test into a conditional two-sample test while preserving the original test's asymptotic properties. The second transforms the problem into comparing marginal distributions via estimated density ratios, enabling the use of established marginal two-sample testing methods. Assuming the distributions have densities, the null hypothesis $f_{Y|X}^{(1)}(y|x) = f_{Y|X}^{(2)}(y|x)$ can be reformulated as

$$f_{XY}^{(2)}(x,y) = \{f_X^{(2)}(x)/f_X^{(1)}(x)\}f_{XY}^{(1)}(x,y), \tag{16}$$

where $f_{XY}^{(1)}$ and $f_{XY}^{(2)}$ are the joint densities. This approach in Lee et al. (2024) aligns with Section 4.1 in that part of the data is used to estimate the marginal density ratio, while the remaining data is used for testing. This is achieved either using a classifier-based test statistic or a linear-time MMD statistic, both reweighted according to the marginal importance, which are proven to result in asymptotically valid tests.

In light of (16), our methodology can be adapted to address the conditional two-sample testing problem. This is achieved by first estimating the marginal density ratio and then applying the DRPT, using this estimate as the shift factor. The key innovation of this approach lies in its calibration via permutations, where the only excess in Type-I error arises from the estimation of the marginal density ratio. A detailed

empirical comparison between this methodology and those proposed in Lee et al. (2024) is provided in Section 5.2.

# 5 Simulations and real-data applications

## 5.1 Synthetic data

In this section, we empirically validate the performance of the DRPT on synthetic data. For the first setting, we set $P_f = N(0, 1)$ and define $P_g = (1 + \eta)^{-1} N(0, \frac{1}{9}) + \eta(1 + \eta)^{-1} \text{Exp}(1)$ for $\eta \in \{0, \ldots, 0.25\}$. Clearly, for $\eta = 0$, $g$ satisfies the null hypothesis with $r(x) = e^{-4x^2}$, while larger values of $\eta$ correspond to greater departures from the null. In the simulations shown by the purple line, we implement the DRPT using the empirical shifted-MMD test statistic defined in (12) combined with the Gaussian kernel $K_{\text{Gauss}}$. The bandwidth of the kernel is calibrated using the median heuristic proposed in Gretton et al. (2012). We compare our approach with a permutation test based on the classical MMD (shown by the green line), applied to the data $Y_1, \ldots, Y_m$ and a sample from $rf$, which is generated from $X_1, \ldots, X_n$ using rejection sampling. Furthermore, we also consider a similar setting (shown by the orange and blue lines) in which $P_f = N(0, 1)$ and $P_g = (1 + \eta)^{-1} N(0, \frac{1}{3}) + \eta(1 + \eta)^{-1} \text{Exp}(1)$, which means that the null hypothesis (1) is satisfied for $r'(x) = e^{-x^2}$. All four simulations are conducted for $\eta \in \{0, \ldots, 0.25\}$, fixing $H = 99$, $n = m = 250$, and $S = 50$. Each test is repeated 500 times for every setting, and the average decision is reported as an estimate of the power function. The results are presented in Figure 1. The DRPT demonstrates superior performance compared to the rejection sampling-based procedure, which is to be expected due to the reduction in effective sample size inherent in such resampling methods. This justifies our choice to forego rejection-sampling schemes and utilise the full sample directly. Furthermore, less extreme shifts, such as $r'$, seem harder to test, as the DRPT shows greater power with more peaked choices like $r$.

Moving beyond univariate settings, Figure 2 shows the performance of the DRPT in the case of bivariate data. Here, we let $P_q$ be an absolutely continuous distribution on $[0, 1]$ with density $q(x) = 2x$. We then choose $P_f = \text{Unif}([0, 1]^2)$ and $P_g = (1 + \eta)^{-1} P_q^{\otimes 2} + \eta(1 + \eta)^{-1} \text{Beta}(\frac{1}{2}, \frac{1}{2})^{\otimes 2}$. This implies that, when $\eta = 0$, $g$ satisfies the null hypothesis with $r(x, y) = 4xy$, while larger values of $\eta$ correspond to greater departures from the null. As before, we implement the DRPT using the U-statistic (12) combined with the Gaussian kernel $K_{\text{Gauss}}$, with the bandwidth calibrated using the median heuristic. The simulation (shown in purple) is conducted for $\eta \in \{0, \ldots, 0.9\}$, fixing $H = 99$, $n = m = 250$, and $S = 50$, and repeating the test 500 times in each setting. We compare this setting with a similar one (shown in orange), where we fix $P_f = \text{Unif}([0, 1]^2)$ and $P_g = (1 + \eta)^{-1} P_q \otimes \text{Unif}([0, 1]) + \eta(1 + \eta)^{-1} \text{Beta}(\frac{1}{2}, \frac{1}{2})^{\otimes 2}$; this implies that the null is satisfied for $r'(x, y) = 2x$ when $\eta = 0$. The results shown in Figure 2 are consistent with the earlier conjecture that shifts that are closer to being constant are more difficult to test.

Table 1 provides empirical evidence for the claim in Section 4.1 that approximating $r_\star$ introduces only a modest inflation in Type-I error, provided the estimator $\hat{r}$ is sufficiently accurate and the test statistic $T$ is appropriately chosen. We consider the setting where $P_f = \text{Unif}([0, 1])$ and $P_g = P_q$, so that the null hypothesis holds with $r_\star(x) = 2x$. Testing sample sizes are fixed at $n = m = 250$, while $r_\star$ is estimated using an independent sample of size $N \in \{100, 200, 1000, 2000\}$, drawn under the null. Estimation is performed using either *linear logistic regression* (LL) (Sugiyama et al., 2010) or the *uLSIF estimator* (Kanamori et al.,

2009). The table reports the average Type-I error across 200 repetitions of the DRPT, using either the true $r_\star$ or an estimate thereof, and the empirical shifted-MMD with the Gaussian kernel as the test statistic. The results indicate that Type-I error inflation is non-negligible when the LL estimator is used for $r_\star$. This is likely due to the parametric nature of the estimator, which is misspecified for the true $r_\star$, resulting in no improvement as the sample size $N$ increases. In contrast, the ULSIF estimator does get closer to $r_\star$ as $N$ grows, and the test size seems to approach the nominal level asymptotically.
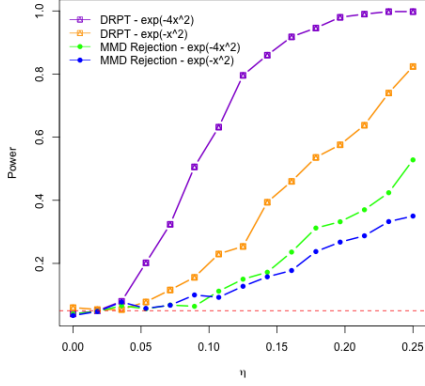


Figure 1: Purple: simulation of the DRPT based on (12) using the Gaussian kernel; here $r(x) = e^{-4x^2}$. Orange: same setting but with $r'(x) = e^{-x^2}$. Green and blue: alternative approaches based on rejection sampling.
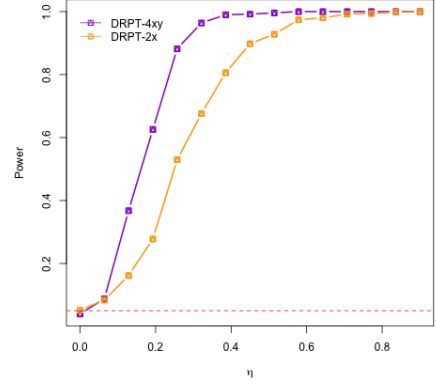


Figure 2: Purple: simulation of the DRPT based on (12) using the Gaussian kernel in the case of bivariate data on the unit square; here $r(x, y) = 4xy$. Orange: same setting but with $r'(x, y) = 2x$.

| Method | $N$ | | | |
|--------|-----|-----|------|------|
|        | 100 | 200 | 1000 | 2000 |
| LL     | 0.520 | 0.250 | 0.065 | 0.140 |
| uLSIF  | 0.095 | 0.055 | 0.085 | 0.065 |
| True   | 0.050 | 0.050 | 0.050 | 0.050 |

Table 1: Type-I error inflation when estimating the density ratio using LL and uLSIF; the true ratio is $r_\star(x) = 2x$; the DRPT was performed using (12) with the Gaussian kernel.
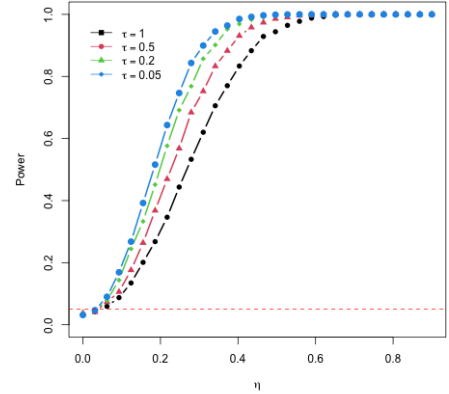


Figure 3: Discrete DRPT based on the approach in Section 2.1 for binary data with varying sample sizes. The test statistic is (11) with the collision kernel.
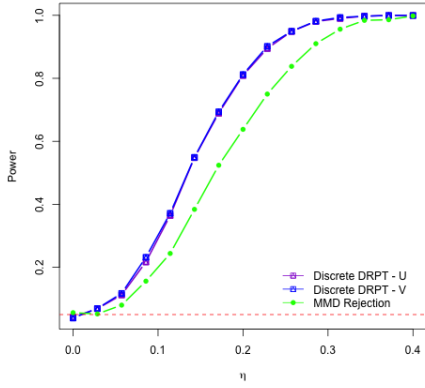
Figure 4: Purple: discrete DRPT using (12) with the collision kernel. Blue: discrete DRPT using (11) with the collision kernel. Green: alternative approach based on rejection sampling. Here $r = (1, 3, 3, 10)$.
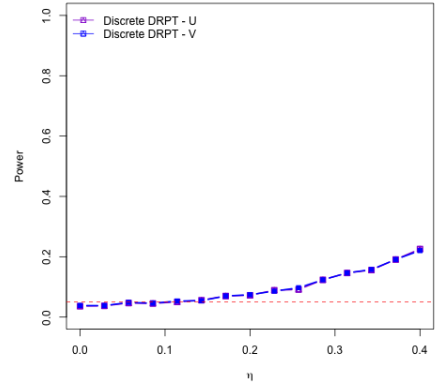


Figure 5: Purple: discrete DRPT using (12) with the collision kernel. Blue: discrete DRPT using (11) with the collision kernel. Here $r = (1, 1, 1, 1)$.

We now turn to discrete settings and evaluate the performance of our RKHS-based methodology, leveraging the discrete DRPT approach introduced in Section 2.1. In this regard, when using the collision kernel $k_{\text{coll}}(x, y) := \sum_{j=0}^{J} \mathbb{1}\{x = j\}\mathbb{1}\{y = j\}$ we can write the V-statistic (11) and the U-statistic (12) as

$$
V(x_1, \ldots, x_n, y_1, \ldots, y_m) = \sum_{j=0}^{J} \frac{1}{(n/m + \widehat{\lambda} r_j)^2} \left\{ \frac{\widehat{\lambda} r_j}{n} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = j\}} - \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{y_i = j\}} \right\}^2,
$$

and

$$
U(x_1, \ldots, x_n, y_1, \ldots, y_m) = V - \sum_{j=0}^{J} \frac{\widehat{\lambda}^2 r_j^2}{(n/m + \widehat{\lambda} r_j)^2} \left\{ \frac{1}{n^2} \sum_{i=1}^{n} \mathbb{1}_{\{x_i = j\}} \right\} - \sum_{j=0}^{J} \frac{1}{(n/m + \widehat{\lambda} r_j)^2} \left\{ \frac{1}{m^2} \sum_{i=1}^{m} \mathbb{1}_{\{y_i = j\}} \right\},
$$

which shows that it is sufficient to draw $(N_{Y,0}^p, \ldots, N_{Y,J}^p)|Z_{()}$ using (7) through the R-function `rMFNCHypergeo` without the need of using Algorithm 1, drastically reducing the computational cost of the DRPT. In the case of binary data, we consider the test statistics (11), fix $n = 100$ and $m \in \{100, 200, 500, 2000\}$, $r = 3$ and sample $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}\left(\frac{1}{2}\right)$ for $i \in [n]$ and $Y_j \overset{\text{i.i.d.}}{\sim} \text{Ber}\left((1 - \eta)\frac{3}{4} + \frac{\eta}{4}\right)$ for $j \in [m]$ with $\eta \in \{0, \ldots, 0.9\}$. The null hypothesis holds for $\eta = 0$, with increasing values of $\eta$ representing greater deviations from the null. We used $H = 99$ and, for each $\eta$, repeated the simulation 5000 times. The results, illustrated in Figure 3, indicate that our methodology performs well even with unbalanced sample sizes, achieving greater power in the most unbalanced case, which corresponds to the largest sample sizes in this setting.

We extended our analysis to a discrete setting with larger finite support. Specifically, we set $J = 3$ and $n = m = 250$, where the $X_i$'s were independently drawn from a multinomial distribution over $\{0, 1, 2, 3\}$ with probabilities $p_X = (\frac{1}{8}, \frac{1}{8}, \frac{3}{8}, \frac{3}{8})$, and the $Y_j$'s were sampled from a multinomial distribution with probabilities $p_Y = (\frac{1}{43}, \frac{3}{43}, \frac{9+25\eta}{43}, \frac{30-25\eta}{43})$, with $\eta \in \{0, \ldots, 0.4\}$. When $\eta = 0$, the null hypothesis

22

holds with $r = (1, 3, 3, 10)$; increasing $\eta$ corresponds to larger deviations from the null. We fixed $S = 50$ and repeated each experiment 5000 times, comparing three different methodologies in this setting. Figure 4 shows the results. The DRPT using the V-statistic (11) and the U-statistic (12) with the collision kernel $k_{\mathrm{coll}}(x, y) := \sum_{j=0}^{J} \mathbb{1}\{x = j\}\mathbb{1}\{y = j\}$ are shown in blue and purple, respectively. The classical MMD permutation test combined with rejection sampling, as described earlier, is shown in green. Overall, the results show a notable drop in power when rejection sampling is used. Furthermore, the tests based on (11) and (12) are nearly indistinguishable, as expected, since the difference $V - U$ vanishes asymptotically as the sample size increases. Finally, Figure 5 presents the analogue of Figure 4 for a related testing problem with a less extreme shift. Specifically, the shift vector is $r = (1, 1, 1, 1)$, so the problem reduces to the classical two-sample testing setting. The $X_i$'s are drawn i.i.d. from the uniform distribution on $\{0, 1, 2, 3\}$, that is, $p_X' = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$, while the $Y_j$'s are sampled from a multinomial distribution with $p_Y' = (\frac{1}{4}, \frac{1}{4}, \frac{1+\gamma}{4}, \frac{1-\gamma}{4})$, where $\gamma = \frac{25\eta}{43}\left(\frac{1}{\sqrt{3}} + \frac{1}{\sqrt{10}}\right)$. This specific choice of $\gamma$ ensures that $D(p_X, p_Y) = D(p_X', p_Y')$ for all values of $\eta$, where the separation $D(f, g)$ is defined in Appendix B. The empirical results in Figure 5 appear to support the earlier conjecture that nearly uniform shifts are, in general, more difficult to detect.

## 5.2 Real-world application

### 5.2.1 Stroop-effect and New-York-frisk datasets

When working with simple classes of distributional shifts, our hypothesis tests can be inverted in the usual way to furnish confidence sets, offering useful insights to practitioners. To illustrate this, we consider two practical scenarios, beginning with the well-known Stroop effect (Stroop, 1935). This refers to the cognitive interference observed when an individual attempts to name the colour of the ink of a word that spells a different colour (incongruent), compared to naming the colour of the ink when the word and ink colour match (congruent). Typically, reaction times and error rates differ significantly between these congruent and incongruent conditions, indicating distinct underlying distributions for concordant (congruent) and non-concordant (incongruent) stimuli. The data set consists of observations from 131 individuals, each with two recorded values: $X$, representing the time taken to name a set of concordant pairs (i.e., colour-word-match) and $Y$, the time taken to name an equal number of discordant pairs (i.e., colour-word mismatched). We standardise the data, and test the hypothesis $g(y) \propto e^{y/\eta} f(y)$ for varying $\eta \in \{0.01, \ldots, 0.3\}$ using the DRPT based on the U-statistic (12) combined with the Gaussian kernel $K_{\mathrm{Gauss}}$. The results in Figure 6 show that there are values for which such modelling assumptions cannot be ruled out, especially in a neighbourhood of 0.2.

In a similar spirit, we also consider the New-York-frisk dataset for the years 2011 and 2012, which contains detailed records of police stop-and-frisk encounters, including demographic, contextual, and outcome-related variables. Although the dataset contains a wide range of features, we limit our analysis to stops recorded as *criminal possession of a weapon* and retain only the indicators reflecting whether different types of weapons were found during the frisk. These indicators are then combined into a single binary variable indicating whether any weapon was found. The sample is then divided into two groups based on whether the individual is identified as Black or White, resulting in two binary datasets: $X$, where a value of 1 indicates weapon possession among Black individuals, and $Y$, where a value of 1 indicates weapon possession among White individuals. The $X$ sample contains 82,626 observations, whereas the $Y$ sample comprises 6,383 observations.
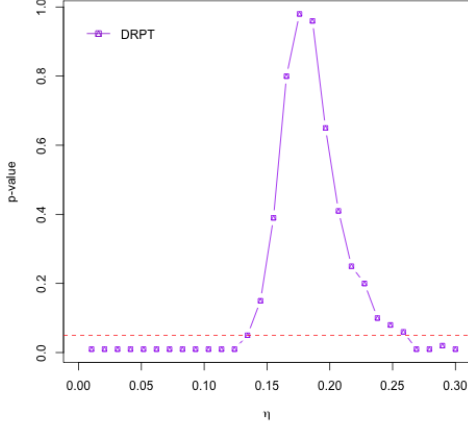
Figure 6: $p$-values of the DRPT testing $g(y) \propto e^{y/\eta} f(y)$ for $\eta \in \{0.01, \ldots, 0.3\}$ on the *Stroop* data. The DRPT was performed using (12) with the Gaussian kernel.
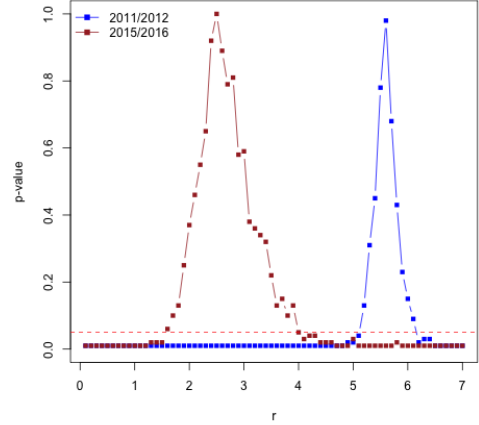
Figure 7: $p$-values of the discrete DRPT testing $w_1/w_0 = rb_1/b_0$ for $r \in \{0.1, \ldots, 7\}$ on the *NY-Frisk* datasets for the years 2011/2012 (blue) and 2015/2016 (brown). The DRPT was performed using (11) with the collision kernel.

A conjecture in the literature, illustrated in Figure 1(b) of Goel et al. (2016) and recalled in Section 8.1 in Koh et al. (2020), suggests that Black individuals are approximately five times less likely to possess a weapon compared to White individuals. We examine this claim using the discrete DRPT using (11) as test statistic. Specifically, when testing the relation $w_1/w_0 = rb_1/b_0$ for $r \in \{0.1, \ldots, 7\}$, where $w_1$ and $b_1$ denote the probabilities of weapon possession for White and Black individuals respectively, the DRPT does not reject the null hypotheses for values of $r$ between 5 and 6. This supports the claim that, conditional on being frisked, Black individuals in 2011–2012 were about five times less likely than White individuals to be found carrying a weapon, even though they were frisked more frequently. We also considered analogous datasets for the years 2015 and 2016, which, after applying the same data cleaning process, contain the records of 138 White individuals and 2,298 Black individuals who were frisked on suspicion of carrying a weapon. The reduction in sample sizes, along with the results in Figure 7, where values of $r$ between 2 and 4 are not rejected, may suggest a change in frisking or stopping practices between 2011 and 2016.

### 5.2.2 Conditional two-sample testing on the diamonds datasets

We now assess the performance of the DRPT in the conditional two-sample testing problem, as outlined in Section 4.2. Our analysis utilises the diamonds dataset, which is available in the R package ggplot2, and contains 53,490 observations with 10 features, including price, carat, clarity, and colour. Following Lee et al. (2024), we designate the price variable as $Y$ and use the six numerical variables (carat, depth, table, x, y, z) as predictors $X$. Prior to the analysis, we standardise both $X$ and $Y$. To introduce covariate shift, we implement biased sampling procedures: $X^{(1)}$ is sampled uniformly from the original feature space, while $X^{(2)}$ is sampled with probabilities proportional to $\exp(-x_1^2)$, with $x_1$ representing the first feature of $X$. Under the null hypothesis, the response variable $Y$ is uniformly sampled for both $Y^{(1)}$ and $Y^{(2)}$. Under the alternative hypothesis, $Y^{(1)}$ remains uniformly sampled, while $Y^{(2)}$ is sampled with probabilities proportional

to $\exp(-y)$, where $y$ corresponds to the dataset's $Y$ values. We compare our methodologies against several hypothesis testing approaches introduced in Lee et al. (2024): the single-split classifier-based test (CLF) and its cross-fit version (†CLF), and the linear-time Maximum Mean Discrepancy test (MMD-$l$) alongside its cross-fit counterpart (†MMD-$l$). Additionally, we include the Conformal Prediction (CP) test based on conformity scores (Hu and Lei, 2020), and the Debiased Conformal Prediction (DCP) test, which enhances CP through Neyman orthogonality and cross-fitting (Chen and Lei, 2025). Results are presented in Table 2, demonstrating that our method achieves effective control of the Type-I error and exhibits competitive power compared to other methodologies. In our simulation, we used an 80/20 split, allocating 80% of the data to marginal density ratio estimation and 20% to testing. Consistently with all the other methods, the marginal density ratio is estimated either with the *linear logistic (LL)* or the *kernel logistic (KLR)* regression. The testing phase was performed using the U-statistic (12) combined with the Gaussian kernel $K_{\mathrm{Gauss}}$.

Table 2: Simulation results for the conditional two-sample testing problem on the diamonds dataset.

| Estimator | Hypothesis | Test | 200 | 400 | 800 | 1200 | 1600 | 2000 |
|---|---|---|---|---|---|---|---|---|
| LL | Null | CLF | 0.0900 | 0.0650 | 0.0750 | 0.0450 | 0.0675 | 0.0425 |
| LL | Null | CP | 0.0650 | 0.0875 | 0.0925 | 0.0575 | 0.1100 | 0.0925 |
| LL | Null | †CLF | 0.0950 | 0.0675 | 0.0850 | 0.0750 | 0.0450 | 0.0675 |
| LL | Null | †MMD-$l$ | 0.0700 | 0.0700 | 0.0675 | 0.0550 | 0.0750 | 0.0625 |
| LL | Null | MMD-$l$ | 0.0750 | 0.0650 | 0.0750 | 0.0500 | 0.0575 | 0.0575 |
| LL | Null | DCP | 0.0375 | 0.0400 | 0.0350 | 0.0425 | 0.0325 | 0.0400 |
| LL | Null | DRPT | 0.0600 | 0.0550 | 0.0850 | 0.0600 | 0.0500 | 0.0350 |
| LL | Alternative | CLF | 0.1575 | 0.2050 | 0.2650 | 0.3600 | 0.3925 | 0.4800 |
| LL | Alternative | CP | 0.2950 | 0.5275 | 0.6900 | 0.8700 | 0.9050 | 0.9300 |
| LL | Alternative | †CLF | 0.2425 | 0.3675 | 0.4700 | 0.6225 | 0.6575 | 0.7575 |
| LL | Alternative | †MMD-$l$ | 0.0975 | 0.1100 | 0.0900 | 0.1075 | 0.1125 | 0.1275 |
| LL | Alternative | MMD-$l$ | 0.0650 | 0.0675 | 0.0850 | 0.0825 | 0.0975 | 0.0850 |
| LL | Alternative | DCP | 0.1750 | 0.4150 | 0.6750 | 0.7925 | 0.8250 | 0.7950 |
| LL | Alternative | DRPT | 0.1500 | 0.1800 | 0.4150 | 0.5450 | 0.6350 | 0.7150 |
| KLR | Null | CLF | 0.0750 | 0.0575 | 0.0675 | 0.0350 | 0.0475 | 0.0450 |
| KLR | Null | CP | 0.0450 | 0.0675 | 0.0575 | 0.0400 | 0.0675 | 0.0500 |
| KLR | Null | †CLF | 0.0825 | 0.0375 | 0.0475 | 0.0400 | 0.0425 | 0.0500 |
| KLR | Null | †MMD-$l$ | 0.0200 | 0.0450 | 0.0675 | 0.0650 | 0.0550 | 0.0550 |
| KLR | Null | MMD-$l$ | 0.0600 | 0.0625 | 0.0725 | 0.0575 | 0.0550 | 0.0600 |
| KLR | Null | DCP | 0.0275 | 0.0150 | 0.0275 | 0.0175 | 0.0275 | 0.0200 |
| KLR | Null | DRPT | 0.0400 | 0.0350 | 0.0600 | 0.0500 | 0.0600 | 0.0300 |
| KLR | Alternative | CLF | 0.0975 | 0.1525 | 0.2600 | 0.3550 | 0.3675 | 0.4450 |
| KLR | Alternative | CP | 0.3100 | 0.5450 | 0.7450 | 0.9025 | 0.9600 | 0.9725 |
| KLR | Alternative | †CLF | 0.1675 | 0.2650 | 0.3900 | 0.5750 | 0.6275 | 0.7150 |
| KLR | Alternative | †MMD-$l$ | 0.0700 | 0.0625 | 0.0950 | 0.1000 | 0.1225 | 0.1425 |
| KLR | Alternative | MMD-$l$ | 0.0725 | 0.0675 | 0.0800 | 0.0900 | 0.0950 | 0.1050 |
| KLR | Alternative | DCP | 0.2425 | 0.4325 | 0.6850 | 0.8175 | 0.9200 | 0.9700 |
| KLR | Alternative | DRPT | 0.1250 | 0.1750 | 0.3750 | 0.5000 | 0.6300 | 0.6650 |

# References

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

Sivaraman Balakrishnan and Larry Wasserman. Hypothesis testing for densities and high-dimensional multinomials: Sharp local minimax rates. *Ann. Statist.*, 47:1893–1927, 2019.

Thomas B. Berrett. Efficient estimation with incomplete data via generalised ANOVA decomposition, 2024. URL https://arxiv.org/abs/2409.05729.

Thomas B Berrett, Yi Wang, Rina Foygel Barber, and Richard J Samworth. The conditional permutation test for independence while controlling for confounders. *J. Roy. Statist. Soc. Ser. B*, 82:175–197, 2020.

Thomas B Berrett, Ioannis Kontoyiannis, and Richard J Samworth. Optimal rates for independence testing via $U$-statistic permutation tests. *Ann. Statist.*, 49:2457–2490, 2021.

Julian Besag and Peter Clifford. Generalized Monte Carlo Significance Tests. *Biometrika*, 76(4):633–642, 1989.

Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, page 81–88, 2007.

Cristina Butucea. Goodness-of-fit testing and quadratic functional estimation from indirect observations. *Annals of Statistics*, 35(5):1907–1930, 2007.

Anirban Chatterjee, Ziang Niu, and Bhaswar B. Bhattacharya. A Kernel-Based Conditional Two-Sample Test Using Nearest Neighbors (with Applications to Calibration, Regression Curves, and Simulation-Based Inference), 2024. URL https://arxiv.org/abs/2407.16550.

Siu Lun Chau, Antonin Schrab, Arthur Gretton, Dino Sejdinovic, and Krikamol Muandet. Credal Two-Sample Tests of Epistemic Ignorance, 2024. URL https://arxiv.org/abs/2410.12921.

Jiahua Chen and Yukun Liu. Quantile and quantile-function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669 – 1692, 2013.

Yuchen Chen and Jing Lei. De-Biased Two-Sample U-Statistics With Application To Conditional Distribution Testing. *Machine Learning*, 114(2):33, 2025.

K.F. Cheng and C.K. Chu. Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583 – 604, 2004.

Kristy Choi, Madeline Liao, and Stefano Ermon. Featurized Density Ratio Estimation. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pages 172–182, 2021.

Kristy Choi, Chenlin Meng, Yang Song, and Stefano Ermon. Density Ratio Estimation via Infinitesimal Classification. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 2552–2573, 2022.

R. M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2 edition, 2002.

P. Economou and G. Tzavelas. Sample Tests for Detection of Size-Biased Sampling Mechanism. *Communications in Statistics - Theory and Methods*, 42(18):3280–3295, 2013.

P. Economou and G. Tzavelas. Kullback–Leibler divergence measure based tests concerning the biasness in a sample. *Statistical Methodology*, 21(Complete):88–108, 2014.

Josh Givens, Song Liu, and Henry W. J. Reeve. Density Ratio Estimation and Neyman Pearson Classification with Missing Data. In *International Conference on Artificial Intelligence and Statistics*, 2023.

Sharad Goel, Justin M. Rao, and Ravi Shroff. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *The Annals of Applied Statistics*, 10(1):365–394, 2016.

A. Gretton, A.J. Smola, J. Huang, Marcel Schmittfull, K.M. Borgwardt, Bernhard Schölkopf, J. Candela, M. Sugiyama, A. Schwaighofer, and N. Lawrence. Covariate Shift by Kernel Mean Matching. *Dataset Shift in Machine Learning, 131-160 (2009)*, 01 2009a.

Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten M. Borgwardt, Bernhard Schölkopf, Quiñonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Covariate Shift by Kernel Mean Matching. In *Neural Information Processing Systems*, 2009b.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Hao Guan and Mingxia Liu. Domain Adaptation for Medical Image Analysis: A Survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2022.

Shohei Hido, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. Statistical Outlier Detection Using Direct Density Ratio Estimation. *Knowledge and Information Systems*, 26:309–336, 02 2011.

Wassily Hoeffding. The Large-Sample Power of Tests Based on Permutations of Observations. *The Annals of Mathematical Statistics*, 23(2):169 – 192, 1952.

D. G. Horvitz and D. J. Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952.

Xiaoyu Hu and Jing Lei. A Two-Sample Conditional Distribution Test Using Conformal Prediction and Weighted Rank Sum. *Journal of the American Statistical Association*, 119:1136 – 1154, 2020.

Ming-Yueh Huang, Jing Qin, and Chiung-Yu Huang. Efficient data integration under prior probability shift. *Biometrics*, 80(2):ujae035, 2024.

Yu. I. Ingster. Minimax Testing of Nonparametric Hypotheses on a Distribution Density in the $l_p$ Metrics. *Theory of Probability and Its Applications*, 31(2):333–337, 1987.

Yuri Ingster and Irina Suslina. *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*, volume 169. Springer Science & Business Media, 2003.

Wenlong Ji, Weizhe Yuan, Emily Getzen, Kyunghyun Cho, Michael I. Jordan, Song Mei, Jason E Weston, Weijie J. Su, Jing Xu, and Linjun Zhang. An Overview of Large Language Models for Statisticians, 2025. URL https://arxiv.org/abs/2502.17814.

Ying Jin and Emmanuel J. Candès. Model-free selective inference under covariate shift via weighted conformal p-values, 2023. URL https://arxiv.org/abs/2307.09291.

Herman Kahn and Theodore E. Harris. Estimation of particle transmission by random sampling. Technical Report 12, National Bureau of Standards, 1951.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A Least-squares Approach to Direct Importance Estimation. *Journal of Machine Learning Research*, 10, 2009.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Theoretical Analysis of Density Ratio Estimation. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 93-A:787–798, 2010a.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. f-Divergence Estimation and Two-Sample Homogeneity Test Under Semiparametric Density-Ratio Models. *IEEE Trans. Inf. Theory*, 58:708–720, 2010b.

Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86:335–367, 2012.

Qing Kang and Paul I. Nelson. Nonparametric tests for the median from a size-biased sample. *Journal of Nonparametric Statistics*, 20:19 – 37, 2008.

Qing Kang and Paul I. Nelson. Permutation tests from biased samples for the equality of two distributions. *Journal of Nonparametric Statistics*, 21(3):305–319, 2009.

Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *Ann. Statist.*, 50(1):225–251, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etiene David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Meghan Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *International Conference on Machine Learning*, 2020.

A. J. Lee. *U-Statistics: Theory and Practice*. Routledge, New York, 1990.

Seongchan Lee, Suman Cha, and Ilmun Kim. General Frameworks for Conditional Two-Sample Testing, 2024. URL https://arxiv.org/abs/2410.16636.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses.* Springer Science & Business Media, 2006.

Tong Li and Ming Yuan. On the Optimality of Gaussian Kernel Based Nonparametric Tests against Smooth Alternatives. *Journal of Machine Learning Research*, 25(334):1–62, 2024.

Song Liu, Akiko Takeda, Taiji Suzuki, and Kenji Fukumizu. Trimmed Density Ratio Estimation. In *Neural Information Processing Systems*, 2017.

Cong Ma, Reese Pathak, and Martin J. Wainwright. Optimally tackling covariate shift in RKHS-based nonparametric regression. *The Annals of Statistics*, 51(2):738 – 761, 2023.

P. McCullagh and J.A. Nelder. *Generalized Linear Models, Second Edition.* Chapman and Hall/CRC Monographs on Statistics and Applied Probability Series. Chapman & Hall, 1989.

Jorge Navarro, José-María Ruiz, and Yolanda del Águila. How to Detect Biased Samples? *Biometrical Journal*, 45(1):91–112, 2003.

XuanLong Nguyen, Martin J. Wainwright, and Michael I. Jordan. Estimating Divergence Functionals and the Likelihood Ratio by Convex Risk Minimization. *IEEE Transactions on Information Theory*, 56:5847–5861, 2008.

Art B. Owen and Yi Zhou. Safe and Effective Importance Sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal Validity Guarantees Exist for Any Data Distribution (and How to Find Them). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 41086–41118, 2024.

Yongsong Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85:619–630, 1998.

Aaditya Ramdas, Rina Foygel Barber, Emmanuel J. Candès, and Ryan J. Tibshirani. Permutation Tests Using Arbitrary Permutation Distributions. *Sankhya A*, 85:1156 – 1177, 2022.

James M. Robins, Miguel A. Hernán, and Babette A. Brumback. Marginal Structural Models and Causal Inference in Epidemiology. *Epidemiology*, 11:550–560, 2000.

Antonin Schrab, Ilmun Kim, Mélisande Albert, Béatrice Laurent, Benjamin Guedj, and Arthur Gretton. MMD Aggregated Two-Sample Test. *Journal of Machine Learning Research*, 24(194):1–81, 2023.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert Space Embeddings and Metrics on Probability Measures. *Journal of Machine Learning Research*, 11:1517–1561, 2010.

Amos Storkey. When Training and Test Sets Are Different: Characterizing Learning Transfer. In *Dataset Shift in Machine Learning*, Neural Information Processing Series, pages 3–28. MIT Press, 2009.

J. Ridley Stroop. Studies of Interference in Serial Verbal Reactions. *Journal of Experimental Psychology*, 18 (6):643–662, 1935.

Masashi Sugiyama and Motoaki Kawanabe. *Machine Learning in Non-Stationary Environments: Introduction to Covariate Shift Adaptation*. The MIT Press, 2012.

Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60:699–746, 2008.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density Ratio Estimation: A Comprehensive Review. *RIMS Kokyuroku*, pages 10–31, 2010.

Masashi Sugiyama, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. Least-squares two-sample test. *Neural networks : the official journal of the International Neural Network Society*, 24 7: 735–51, 2011.

Masashi Sugiyama, Teruyuki Suzuki, and Takafumi Kanamori. Density-ratio matching under the Bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64:1009–1044, 2012.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. A Bradford Book, 2018.

Nikolaj Thams, Sorawit Saengkyongam, Niklas Pfister, and Jonas Peters. Statistical testing under distributional shifts. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(3):597–663, 2023.

Ryan J Tibshirani, Rina Foygel Barber, Emmanuel Candes, and Aaditya Ramdas. Conformal Prediction Under Covariate Shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Surya T. Tokdar and Robert E. Kass. Importance Sampling: A Review. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(1):54–60, 2010.

Yuta Tsuboi, H. Kashima, S. Hido, Steffen Bickel, and M. Sugiyama. Direct Density Ratio Estimation for Large-scale Covariate Shift Adaptation. *Journal of Information Processing*, 17:138–155, 01 2009.

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

Karl Weiss, Taghi M. Khoshgoftaar, and Dingding Wang. A Survey of Transfer Learning. *Journal of Big Data*, 3(1):9, 2016.

Hsin wen Chang and Shu-Hsiang Wang. Bivariate Analysis of Distribution Functions Under Biased Sampling. *The American Statistician*, 78:171 – 179, 2023.

Makoto Yamada and Masashi Sugiyama. Dependence Minimizing Regression with Model Selection for Non-Linear Causal Inference under Non-Gaussian Noise. *Proceedings of the AAAI Conference on Artificial Intelligence*, 24(1):643–648, 2010.

Jian Yan, Zhuoxi Li, and Xianyang Zhang. Distance and Kernel-Based Measures for Global and Local Two-Sample Conditional Distribution Testing, 2024. URL https://arxiv.org/abs/2210.08149.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. On Function Approximation in Reinforcement Learning: Optimism in the Face of Large State Spaces. In *Advances in Neural Information Processing Systems*, volume 33, pages 13903–13916, 2020.

# Appendices

Appendix A provides the proofs for all results stated in the main text, while Appendix B contains additional results about the power of the DRPT in the discrete, finite-support setting introduced in Subsection 2.1.

## Appendix A   Proofs

### A.1   Proofs for Section 2

*Proof of Theorem 1.* From the discussion that led to Equation (4) we know that, under $H_0$, the true data vector $Z$ and the DRPT copies $Z^{(1)}, \ldots, Z^{(H)}$ are permutations of $Z_{()}$ obtained via i.i.d. draws from (4), conditional on $Z_{()}$. Therefore, after marginalising over $Z_{()}$, the $H+1$ random variables $\left(Z, Z^{(1)}, \ldots, Z^{(H)}\right)$ are exchangeable. This is sufficient to prove finite-sample validity for every test statistic $T$. □

*Proof of Proposition 2.* The proof consists of simply checking the detailed balance equations for the Markov chain defined by the algorithm. Denote with $\mathcal{P}$ the collection of all $K$ couples of indices $\{(i_1, j_1), \ldots, (i_K, j_K)\}$ such that $(i_1, \ldots, i_K)$ contains distinct elements from $[n]$, and $(j_1, \ldots, j_K)$ contains distinct elements from $\{n+1, \ldots, n+m\}$. For any $\tau \in \mathcal{P}$ and any permutations $p, p'$, we write $p \sim_\tau p'$ if $p$ can be transformed to $p'$ by swapping any subset of the pairs in $\tau$. We now compute the transition probability matrix of the Markov chain defined by Algorithm 1. Every probability sign has to be intended conditionally on $Z_{()}$. For all $t \in \mathbb{N}_+$ and any permutations $p, p'$, we have

$$\mathbb{P}\left\{P_t = p' \mid P_{t-1} = p\right\} = \frac{1}{|\mathcal{P}|} \sum_{\tau \in \mathcal{P}} \mathbb{P}\left\{P_t = p' \mid P_{t-1} = p, \tau_t = \tau\right\},$$

since at each time $t$, Step 3 of the algorithm corresponds to drawing $\tau_t \in \mathcal{P}$ uniformly at random. Next, given $\tau_t = \tau := \{(i_1, j_1), \ldots, (i_K, j_K)\}$ and $P_{t-1} = p$, it must be the case that $P_t$ satisfies $P_t \sim_\tau p$ by definition of Steps 4-5 of the algorithm. In light of the definition of the odds ratio for each $B_{i_k, j_k}^t$ in (5), we see that for any $p', p'' \sim_\tau p$, we have

$$\frac{\mathbb{P}\left\{P_t = p' \mid P_{t-1} = p, \tau_t = \tau\right\}}{\mathbb{P}\left\{P_t = p'' \mid P_{t-1} = p, \tau_t = \tau\right\}} = \prod_{j \in \{j_1, \ldots, j_K\}} \frac{r(Z_{(p'(j))})}{r(Z_{(p''(j))})} = \prod_{j \in \{j_1, \ldots, j_K\}} \frac{r(Z_{(p'(j))})}{r(Z_{(p''(j))})} \prod_{j \notin \{j_1, \ldots, j_K\}} \frac{r(Z_{(p'(j))})}{r(Z_{(p''(j))})}$$

$$= \prod_{j \in \{n+1, \ldots, n+m\}} \frac{r(Z_{(p'(j))})}{r(Z_{(p''(j))})} = \frac{\mathbb{P}\left\{P = p'\right\}}{\mathbb{P}\left\{P = p''\right\}}, \tag{17}$$

where in the second equality we used the fact that $r(Z_{(p'(j))})/r(Z_{(p''(j))}) = 1$ for all $j \notin \{j_1, \ldots, j_K\}$, while in the last step we used the definition of the distribution (4) conditional on $Z_{()}$. Therefore,

$$\mathbb{P}\{P_t = p' \mid P_{t-1} = p\} = \frac{1}{|\mathcal{P}|} \sum_{\tau \in \mathcal{P}} \frac{\mathbb{1}\{p' \sim_\tau p\} \cdot \mathbb{P}\{P = p'\}}{\sum_{p''} \mathbb{1}\{p'' \sim_\tau p\} \cdot \mathbb{P}\{P = p''\}}.$$

This, together with the fact that $\sim_\tau$ defines an equivalence relation on $\mathcal{P}$, shows that

$$\mathbb{P}\{P = p\} \cdot \mathbb{P}\{P_t = p' \mid P_{t-1} = p\} = \frac{1}{|\mathcal{P}|} \sum_{\tau \in \mathcal{P}} \mathbb{P}\{P = p\} \cdot \frac{\mathbb{1}\{p' \sim_\tau p\} \cdot \mathbb{P}\{P = p'\}}{\sum_{p''} \mathbb{1}\{p'' \sim_\tau p\} \cdot \mathbb{P}\{P = p''\}}$$

$$= \frac{1}{|\mathcal{P}|} \sum_{\tau \in \mathcal{P}} \mathbb{P}\{P = p'\} \cdot \frac{\mathbb{1}\{p \sim_\tau p'\} \cdot \mathbb{P}\{P = p\}}{\sum_{p''} \mathbb{1}\{p'' \sim_\tau p'\} \cdot \mathbb{P}\{P = p''\}} = \mathbb{P}\{P = p'\} \cdot \mathbb{P}\{P_t = p \mid P_{t-1} = p'\}. \qquad (18)$$

This verifies the detailed balance equations, and so the Markov chain is reversible and has (4) as stationary distribution. Moreover, since $r(x)$ is assumed to be positive for all $x \in \mathcal{X}$, it follows that the chain is aperiodic and irreducible, ensuring the uniqueness of the stationary distribution. $\qquad \square$

*Proof of Proposition 3.* The proof relies fundamentally on the reversibility of the Markov chain defined in Algorithm 1, a property established in the proof of Proposition 2. This reversibility permits an alternative but equivalent sampling procedure under the null hypothesis: we may first sample $P_*$ from distribution (4) conditional on $Z_{()}$, and subsequently generate $(P, P^{(1)}, \ldots, P^{(H)})$ through $H + 1$ independent applications of Algorithm 1, each running for $S$ steps and initialised at $P_*$. The independence of these runs, combined with their shared initialisation point $P_*$, ensures that $(P, P^{(1)}, \ldots, P^{(H)})$ are independently and identically distributed when conditioned on $P_*$ and $Z_{()}$. This conditional independence directly implies their exchangeability, and concludes the proof. $\qquad \square$

## A.2 Proofs for Section 3

*Proof of Lemma 4.* We begin by proving that there exists a unique $\lambda_0$ such that $h = \frac{nf + mg}{n + \lambda_0 mr}$ is a density function. Define the function

$$F : \begin{cases} \mathbb{R}_+ \to \mathbb{R}_+ \\ \lambda \mapsto \int \frac{nf + mg}{n + \lambda mr} d\mu \end{cases} \qquad (19)$$

which can easily be seen to be continuous. It also straightforward to see that $F$ is strictly decreasing, since if $\lambda_2 > \lambda_1 > 0$ we have

$$\frac{nf(x) + mg(x)}{n + \lambda_2 mr(x)} \leq \frac{nf(x) + mg(x)}{n + \lambda_1 mr(x)} \quad \text{for all } x \in \mathcal{X},$$

and this inequality is strict in the support of $f$ and $g$ as we are assuming that $r(x) > 0$ for all $x \in \mathcal{X}$. It is clear that $\lim_{\lambda \to 0} F(\lambda) = 1 + m/n > 1$, while $\lim_{\lambda \to +\infty} F(\lambda) = 0 < 1$. Thus, by the intermediate value theorem and the strict monotonicity of $F$ we have established our first claim that $\lambda_0$ exists and is unique.

It can now be seen that $h$ satisfies the requirements of the result. Indeed, we have

$$1 = \int h d\mu = \int \frac{nf + mg}{n + \lambda_0 mr} d\mu = \frac{1}{n} \int \frac{n}{n + \lambda_0 mr} (nf + mg) d\mu$$

32

$$= \frac{1}{n} \int \left( 1 - \frac{\lambda_0 mr}{n + \lambda_0 mr} \right) (nf + mg) d\mu = \frac{1}{n} \left( \int (nf + mg) d\mu - \int (\lambda_0 mr) \frac{nf + mg}{n + \lambda_0 mr} d\mu \right)$$

$$= \frac{1}{n} \left( n + m - \lambda_0 m \int rh d\mu \right) = 1 + \frac{m}{n} - \lambda_0 \frac{m}{n} \int rh d\mu,$$

which implies that $\lambda_0 = (\int rh d\mu)^{-1}$. This shows us that $nh + m(\int rh d\mu)^{-1} rh = nh + \lambda_0 mrh = nf + mg$, so that $h$ gives rise to null distributions for which the distribution of the combined sample matches that of our data, concluding the existence part of the result.

As for its uniqueness, observe that any density $h$ preserving the combined distribution under the null hypothesis must be of the form $h$ for some $\lambda_0$ positive, since

$$nf + mg = nh + m \frac{rh}{\int rf d\mu} = \left\{ n + mr \left( \int rh d\mu \right)^{-1} \right\} h = (n + \lambda_0 mr) h.$$

But the $\lambda_0$ such that $h$ integrates to 1 is unique. This completes the proof.

$\square$

*Proof of Lemma 5.* Suppose that the null hypothesis holds, so that $g = rf / \int rf d\mu$. It follows from Lemma 4 and its proof that $f = h$ and that $\lambda_0 = (\int rf d\mu)^{-1}$. We therefore see that $\lambda_0 rf - g = 0$, so indeed $T_{\mathcal{F},r}(f, g) = 0$.

We now turn to the reverse implication. Assuming that $f$ and $g$ are such that $T_{\mathcal{F},r}(f, g) = 0$, we will show that $H_0$ must be true. By assumption, $\mathcal{H}$ is dense in $C_b^0(\mathcal{X})$ with respect to $\| \cdot \|_\infty$, so that for all $\phi \in C_b^0(\mathcal{X})$ and all $\epsilon > 0$ there exists $\varphi \in \mathcal{H}$ such that $\|\varphi - \phi\|_\infty \leq \epsilon$. Thus

$$\left| \int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r} \phi \, d\mu \right| \leq \left| \int \frac{\lambda_0 rf}{n/m + \lambda_0 r} (\phi - \varphi) d\mu \right| + \left| \int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r} \varphi d\mu \right| + \left| \int \frac{g}{n/m + \lambda_0 r} (\phi - \varphi) d\mu \right|$$

$$\leq \int f|\phi - \varphi| d\mu + \|\varphi\|_{\mathcal{H}} \, T_{\mathcal{F},r}(f, g) + \frac{m}{n} \int g|\phi - \varphi| d\mu$$

$$= \int f|\phi - \varphi| d\mu + \frac{m}{n} \int g|\phi - \varphi| d\mu \leq (1 + m/n)\epsilon,$$

where in the only equality we used our assumption that $T_{\mathcal{F},r}(f, g) = 0$. As $\epsilon > 0$ was arbitrary, we now see that $\int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r} \phi \, d\mu = 0$ for all $\phi \in C_b^0(\mathcal{X})$, which further implies

$$\frac{mg}{n + \lambda_0 mr} = \frac{\lambda_0 mrf}{n + \lambda_0 mr} \quad \text{a.s.}$$

by Lemma 9.3.2 in Dudley (2002). Simplifying this inequality we see that $g = \lambda_0 rf \propto rf$, as required. $\square$

*Proof of Theorem 6.* We assume throughout the proof that $H_0$ does not hold and aim to show that

$$\mathbb{P}\{\text{DRPT does not reject } H_0\} \longrightarrow 0$$

as $n \to \infty$. Since we assume that $H > \lceil 1/\alpha - 1 \rceil$ we have $\alpha(1 + H) - 1 > 0$ and we can therefore apply

Markov's inequality to see that

$$\mathbb{P}\{\text{DRPT does not reject } H_0\} = \mathbb{P}\left\{1 + \sum_{h=1}^{H} \mathbb{1}\{T(Z_{\sigma^{(h)}}) \geq T(Z)\} > \alpha(1+H)\right\}$$

$$\leq \frac{\mathbb{E}\left[\sum_{h=1}^{H} \mathbb{1}\{T(Z_{\sigma^{(h)}}) \geq T(Z)\}\right]}{\alpha(1+H) - 1} = \frac{H}{\alpha(1+H) - 1}\mathbb{P}\{T(Z_{\sigma^{(1)}}) \geq T(Z)\},$$

where the last step follows from exchangeability. Using the shorthand $\sigma = \sigma^{(1)}$ it now suffices to show that there exists $\eta > 0$ such that $T(Z) \xrightarrow{\mathbb{P}} \eta$ and $T(Z_\sigma) \xrightarrow{\mathbb{P}} 0$.

We will first prove that $\mathbb{E}[|T(Z_\sigma)|] = \mathbb{E}[T(Z_\sigma)] \longrightarrow 0$. Write

$$T_\varphi^\sigma := \frac{1}{n}\left(\sum_{i=1}^{n} \frac{\widehat{\lambda}mr(Z_{\sigma(i)})}{n + \widehat{\lambda}mr(Z_{\sigma(i)})}\varphi(Z_{\sigma(i)}) - \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda}mr(Z_{\sigma(j)})}\varphi(Z_{\sigma(j)})\right),$$

so that we have $T(Z_\sigma) = \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} |T_\varphi^\sigma|$. Write $N \equiv N(\delta) \equiv N(\{\|\varphi\|_{\mathcal{H}} \leq 1\}, \delta, \|\cdot\|_\infty)$ for the covering number of $\{\|\varphi\|_{\mathcal{H}} \leq 1\}$ with respect to $\|\cdot\|_\infty$ and let $\{\psi_1, \ldots, \psi_N\}$ be an associated $\delta$-cover. Now, for a generic function $\varphi_0 \in \mathcal{H}$ such that $\|\varphi_0\|_{\mathcal{H}} \leq 1$ we have

$$\mathbb{E}[T(Z_\sigma)] = \mathbb{E}\left[\sup_{\|\varphi\|_{\mathcal{H}} \leq 1}(|T_\varphi^\sigma| - |T_{\varphi_0}^\sigma|)\right] + \mathbb{E}[|T_{\varphi_0}^\sigma|]$$

$$\leq \mathbb{E}\left[\sup_{\substack{\|\varphi_1\|_{\mathcal{H}} \leq 1 \\ \|\varphi_2\|_{\mathcal{H}} \leq 1}}(|T_{\varphi_1}^\sigma| - |T_{\varphi_2}^\sigma|)\right] + \mathbb{E}[|T_{\varphi_0}^\sigma|] \leq \mathbb{E}\left[\sup_{\substack{\|\varphi_1\|_{\mathcal{H}} \leq 1 \\ \|\varphi_2\|_{\mathcal{H}} \leq 1}}\left|T_{\varphi_1}^\sigma - T_{\varphi_2}^\sigma\right|\right] + \mathbb{E}[|T_{\varphi_0}^\sigma|]$$

$$\leq 2\mathbb{E}\left[\sup_{\substack{\|\varphi_1\|_{\mathcal{H}} \leq 1 \\ \|\varphi_2\|_{\mathcal{H}} \leq 1 \\ \|\varphi_1 - \varphi_2\|_\infty \leq \delta}}\left|T_{\varphi_1}^\sigma - T_{\varphi_2}^\sigma\right|\right] + 2\mathbb{E}\left[\max_{i \in [N]}\left|T_{\psi_1}^\sigma - T_{\psi_i}^\sigma\right|\right] + \mathbb{E}[|T_{\varphi_0}^\sigma|]$$

$$\leq 2\mathbb{E}\left[\sup_{\|\varphi\|_\infty \leq \delta}|T_\varphi^\sigma|\right] + (4N + 1)\sup_{\|\varphi\|_{\mathcal{H}} \leq 1}\mathbb{E}[|T_\varphi^\sigma|], \tag{20}$$

where the penultimate inequality follows from a one-step discretisation argument as in Equation 5.34 in Wainwright (2019) and the final inequality follows on bounding the maximum by a sum, using the fact that $T_{\varphi_1}^\sigma - T_{\varphi_2}^\sigma = T_{\varphi_1 - \varphi_2}^\sigma$ for any $\varphi_1, \varphi_2$, and using the triangle inequality to say that $\max_{i \in [N]}\|\psi_1 - \psi_i\|_{\mathcal{H}} \leq 2$. We will now bound each term on the right-hand side of (20) separately. As for the first, observe that

$$|T_\varphi^\sigma| = \frac{1}{n}\left|\sum_{i=1}^{n} \frac{\widehat{\lambda}mr(Z_{\sigma(i)})}{n + \widehat{\lambda}mr(Z_{\sigma(i)})}\varphi(Z_{\sigma(i)}) - \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda}mr(Z_{\sigma(j)})}\varphi(Z_{\sigma(j)})\right|$$

$$\leq \frac{1}{n}\left\{\sum_{i=1}^{n}\left|\frac{\widehat{\lambda}mr(Z_{\sigma(i)})}{n + \widehat{\lambda}mr(Z_{\sigma(i)})}\varphi(Z_{\sigma(i)})\right| + \sum_{j=n+1}^{n+m}\left|\frac{n}{n + \widehat{\lambda}mr(Z_{\sigma(j)})}\varphi(Z_{\sigma(j)})\right|\right\} \leq \frac{n+m}{n}\|\varphi\|_\infty$$

for any $\varphi \in \mathcal{H}$, which implies that

$$\mathbb{E}\left[\sup_{\|\varphi\|_\infty \leq \delta} |T_\varphi^\sigma|\right] \leq \frac{n+m}{n}\delta.$$

We now turn to the second term in (20). With our assumption that $\|\cdot\|_\infty \leq \gamma\|\cdot\|_\mathcal{H}$ for a constant $\gamma$, it follows from the Cauchy–Schwarz inequality, Equation (25) and Lemma 7 $(i)$ that for any $\varphi$ such that $\|\varphi\|_\mathcal{H} \leq 1$ we have

$$\mathbb{E}[|T_{\varphi_0}^\sigma|] \leq \left\{\mathbb{E}\left[(T_\varphi^\sigma)^2\right]\right\}^{1/2} \leq \gamma\sqrt{\frac{(c+C)}{c}\frac{n+m}{n^2}}.$$

Putting these bounds together, we see that for any $\delta > 0$ we have

$$\mathbb{E}[T(Z_\sigma)] \leq 2\frac{n+m}{n}\delta + \{4N(\delta) + 1\}\gamma\sqrt{\frac{(c+C)}{c}\frac{n+m}{n^2}}.$$

Since we assume that $N\left(\{\|\varphi\|_\mathcal{H} \leq 1\}, \delta, \|\cdot\|_\infty\right)$ is finite for all $\delta > 0$, there exists a sequence $(\delta_n)_{n\in\mathbb{N}_+}$ such that $\delta_n \searrow 0$ and $N\left(\{\|\varphi\|_\mathcal{H} \leq 1\}, \delta_n, \|\cdot\|_\infty\right) \leq n^{1/4}$, for all $n \in \mathbb{N}_+$. This implies that $\delta_n + n^{-1/2}N(\delta_n) \to 0$, so that $\mathbb{E}[T(Z_\sigma)] \to 0$, and we therefore have that $T(Z_\sigma) \xrightarrow{\mathbb{P}} 0$.

It now remains to study the behaviour of the unpermuted statistic $T(Z)$. Define the population quantity

$$T_{\mathcal{F},r} = \sup_{\|\varphi\|_\mathcal{H} \leq 1} |I_\varphi|, \quad \text{where } I_\varphi := \int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r}\varphi \, d\mu$$

with $\lambda_0$ such that $\int \frac{nf+mg}{n+\lambda_0 mr}d\mu = 1$. Furthermore, define its asymptotic counterpart to be

$$T_{\mathcal{F},r}^\infty = \sup_{\|\varphi\|_\mathcal{H} \leq 1} |I_\varphi^\infty|, \quad \text{where } I_\varphi^\infty := \int \frac{\lambda_\infty rf - g}{\tau + \lambda_\infty r}\varphi \, d\mu$$

with $\lambda_\infty$ such that $\int \frac{\tau f+g}{\tau+\lambda_\infty r}d\mu = 1$. As we are working under the alternative hypothesis, we know by Lemma 5 that $T_{\mathcal{F},r}^\infty > 0$. Writing id for the identity permutation, we have

$$|T(Z) - T_{\mathcal{F},r}| = \left|\sup_{\|\varphi\|_\mathcal{H} \leq 1}|T_\varphi^{\text{id}}| - \sup_{\|\varphi\|_\mathcal{H} \leq 1}|I_\varphi|\right| \leq \left|\sup_{\|\varphi\|_\mathcal{H} \leq 1}(|T_\varphi^{\text{id}}| - |I_\varphi|)\right| \leq \sup_{\|\varphi\|_\mathcal{H} \leq 1}\left|T_\varphi^{\text{id}} - I_\varphi\right|$$

$$= \sup_{\|\varphi\|_\mathcal{H} \leq 1}\left|\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\lambda}mr(X_i)}{n + \widehat{\lambda}mr(X_i)}\varphi(X_i) - \frac{1}{n}\sum_{j=1}^m \frac{n}{n + \widehat{\lambda}mr(Y_j)}\varphi(Y_j) - \int \frac{\lambda_0 rf - g}{n/m + \lambda_0 r}\varphi d\mu\right|$$

$$\leq \sup_{\|\varphi\|_\mathcal{H} \leq 1}\left|\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\lambda}mr(X_i)}{n + \widehat{\lambda}mr(X_i)}\varphi(X_i) - \int \frac{\lambda_0 mrf}{n + \lambda_0 mr}\varphi \, d\mu\right|$$

$$+ \sup_{\|\varphi\|_\mathcal{H} \leq 1}\left|\frac{1}{m}\sum_{j=1}^m \frac{m}{n + \widehat{\lambda}mr(Y_j)}\varphi(Y_j) - \int \frac{mg}{n + \lambda_0 mr}\varphi \, d\mu\right|.$$

These two terms can be bounded by almost identical arguments so we will restrict attention to the first. By

35

the triangle inequality this is bounded by

$$\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\widehat{\lambda}mr(X_i)}{n+\widehat{\lambda}mr(X_i)}\varphi(X_i)-\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\varphi(X_i)\right|$$

$$+\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\varphi(X_i)-\int\frac{\lambda_0 mrf}{n+\lambda_0 mr}\varphi d\mu\right|$$

$$\leq|\widehat{\lambda}/\lambda_0-1|\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\frac{n}{n+\widehat{\lambda}mr(X_i)}|\varphi(X_i)|\right\}$$

$$+\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\varphi(X_i)-\mathbb{E}\left[\frac{\lambda_0 mr(X_1)}{n+\lambda_0 mr(X_1)}\varphi(X_1)\right]\right|$$

$$\leq\gamma|\widehat{\lambda}/\lambda_0-1|+\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\varphi(X_i)-\mathbb{E}\left[\frac{\lambda_0 mr(X_1)}{n+\lambda_0 mr(X_1)}\varphi(X_1)\right]\right|,$$

where in the last inequality we used our assumption that $\|\cdot\|_{\infty}\leq\gamma\|\cdot\|_{\mathcal{H}}$. By Lemma 12 below and the fact that $C^{-1}\leq\lambda_0\leq c^{-1}$, we have $\mathbb{E}|\widehat{\lambda}/\lambda_0-1|=O(n^{-1/2})$ as $n\to\infty$. Turning to the second term, and recalling that we write $N\equiv N(\delta)\equiv N(\{\|\varphi\|_{\mathcal{H}}\leq 1\},\delta,\|\cdot\|_{\infty})$ for the covering number of $\{\|\varphi\|_{\mathcal{H}}\leq 1\}$ with respect to $\|\cdot\|_{\infty}$ and $\{\psi_1,\ldots,\psi_N\}$ for an associated $\delta$-cover, we have

$$\mathbb{E}\left[\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\varphi(X_i)-\mathbb{E}\left\{\frac{\lambda_0 mr(X_1)}{n+\lambda_0 mr(X_1)}\varphi(X_1)\right\}\right|\right]$$

$$\leq 2\delta+\mathbb{E}\left[\max_{j\in[N]}\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\psi_j(X_i)-\mathbb{E}\left\{\frac{\lambda_0 mr(X_1)}{n+\lambda_0 mr(X_1)}\psi_j(X_1)\right\}\right|\right]$$

$$\leq 2\delta+N\max_{j\in[N]}\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\frac{\lambda_0 mr(X_i)}{n+\lambda_0 mr(X_i)}\psi_j(X_i)-\mathbb{E}\left\{\frac{\lambda_0 mr(X_1)}{n+\lambda_0 mr(X_1)}\psi_j(X_1)\right\}\right|\right]\leq 2\delta+\frac{\gamma N(\delta)}{n^{1/2}}$$

for any $\delta>0$. As earlier in the proof, we can choose an appropriate sequence $(\delta_n)$ such that this right-hand side converges to zero as $n\to\infty$. Combining our previous bounds, we see that

$$\mathbb{E}|T(Z)-T_{\mathcal{F},r}|\to 0. \tag{21}$$

Furthermore, we have that

$$|T_{\mathcal{F},r}-T_{\mathcal{F},r}^{\infty}|=\left|\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}|I_\varphi|-\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}|I_\varphi^{\infty}|\right|\leq\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}|I_\varphi-I_\varphi^{\infty}|$$

$$=\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\left|\int\left(\frac{\lambda_0 rf-g}{n/m+\lambda_0 r}-\frac{\lambda_\infty rf-g}{\tau+\lambda_\infty r}\right)\varphi\,d\mu\right|$$

$$\leq\left(\sup_{\|\varphi\|_{\mathcal{H}}\leq 1}\|\varphi\|_{\infty}\right)\int\left|\frac{\lambda_0 rf-g}{n/m+\lambda_0 r}-\frac{\lambda_\infty rf-g}{\tau+\lambda_\infty r}\right|d\mu\leq\gamma\int\left|\frac{\lambda_0 rf-g}{n/m+\lambda_0 r}-\frac{\lambda_\infty rf-g}{\tau+\lambda_\infty r}\right|d\mu$$

$$\leq\gamma\int\frac{|\tau-n/m|+|\lambda_\infty-\lambda_0|r}{(n/m+\lambda_0 r)(\tau+\lambda_\infty r)}(\lambda_\infty rf+g)d\mu+\gamma\int\frac{|\lambda_\infty-\lambda_0|}{n/m+\lambda_0 r}rfd\mu$$

$$\leq\gamma\{1+m/n\}\{|\tau m/n-1|+|\lambda_\infty/\lambda_0-1|\}+\gamma|\lambda_\infty/\lambda_0-1|. \tag{22}$$

36

Since $\lambda_0, \lambda_\infty \in [C^{-1}, c^{-1}]$ under $0 < c \le r(\cdot) \le C$, we can show that the above expression converges to zero provided that $n/m \to \tau$ and $\lambda_0 \to \lambda_\infty$. To justify the latter claim, it suffices to observe that (19) converges uniformly over $[C^{-1}, c^{-1}]$ to

$$F^\infty : \begin{cases} \mathbb{R}_+ \to \mathbb{R}_+ \\ \lambda \mapsto \int \dfrac{\tau f + g}{\tau + \lambda r} \, d\mu \end{cases}$$

as $n, m \to \infty$ by the dominated convergence theorem. This, together with the continuity and the strict monotonicity of $F^\infty$ and the functions in (19), implies the pointwise convergence of the inverse functions, thus showing $\lambda_0 \to \lambda_\infty$. Combining (21) and (22), we obtain

$$\mathbb{E}\left|T(Z) - T^\infty_{\mathcal{F}, r}\right| \le \mathbb{E}\left|T(Z) - T_{\mathcal{F}, r}\right| + \left|T_{\mathcal{F}, r} - T^\infty_{\mathcal{F}, r}\right| \to 0,$$

which yields $T(Z) \xrightarrow{\mathbb{P}} T^\infty_{\mathcal{F}, r} > 0$, thereby completing the proof.

$\square$

**Lemma 12.** *Assume $m \le n \le \tau m$ for $\tau \ge 1$, and $0 < c \le r(x) \le C$ for all $x \in \mathcal{X}$. Let $\widehat{\lambda}$ be such that*

$$\sum_{i=1}^n \frac{\widehat{\lambda} m r(X_i)}{n + \widehat{\lambda} m r(X_i)} = \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r(Y_j)}$$

*and $\lambda_0$ such that*

$$\int \frac{nf + mg}{n + \lambda_0 m r} \, d\mu = 1.$$

*There exists a constant $Q_0 \equiv Q_0(p, c, C, \tau) > 0$ such that*

$$\mathbb{E}[|\widehat{\lambda} - \lambda_0|^p] \le Q_0 \, n^{-p/2} \text{ for all } p \in \mathbb{N}.$$

*Proof.* We know that $\widehat{\lambda}$ is a Z-estimator, being the solution with respect to $\lambda$ of

$$\Psi_{n,m}(\lambda) := \frac{1}{n+m} \sum_{k=1}^{n+m} \psi_\lambda(Z_k) = 0, \quad \text{with} \quad \psi_\lambda(x) := \frac{n+m}{n + \lambda m r(x)} - 1,$$

whereas $\lambda_0$ is a population quantity and satisfies

$$\int \frac{nf}{n + \lambda_0 m r} d\mu + \int \frac{mg}{n + \lambda_0 m r} d\mu = 1. \tag{23}$$

We can use the assumptions $m \le n \le \tau m$ and $0 < c \le r(\cdot) \le C$ to show that

$$\frac{\partial \psi_\lambda(x)}{\partial \lambda} = -\frac{(n+m)mr(x)}{\{n + \lambda m r(x)\}^2} \le -\frac{c\tau^{-1}(1 + \tau^{-1})n^2}{(1 + C/c)^2 n^2} = -\frac{c^3(1+\tau)}{\tau^2(c+C)^2} =: -a(c, C, \tau) \equiv -a < 0,$$

which implies that for all $\epsilon > 0$ we have $\{|\widehat{\lambda} - \lambda_0| < \epsilon\} \supseteq \{|\Psi_{n,m}(\widehat{\lambda}) - \Psi_{n,m}(\lambda_0)| < a\epsilon\} = \{|\Psi_{n,m}(\lambda_0)| < a\epsilon\}$, using the fact that $\Psi_{n,m}(\widehat{\lambda}) = 0$ by definition of $\widehat{\lambda}$. As a result, the previous inclusion shows that the boundedness assumption on $r(\cdot)$ allows to relate how close $\widehat{\lambda}$ is to $\lambda_0$ with $|\Psi_{n,m}(\lambda_0)|$, which is easier to

analyse since it is the sum of zero mean, bounded i.i.d. random variables. In this regard, we have

$$\mathbb{P}\{|\widehat{\lambda} - \lambda_0| \geq \epsilon\} \leq \mathbb{P}\{|\Psi_{n,m}(\lambda_0)| \geq a\epsilon\} = \mathbb{P}\left\{\left|\frac{1}{n+m}\sum_{k=1}^{n+m}\left(\frac{n+m}{n+\lambda_0 mr(Z_k)} - 1\right)\right| \geq a\epsilon\right\}$$

$$= \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n+\lambda_0 mr(X_i)} + \frac{1}{m}\sum_{j=1}^{m}\frac{m}{n+\lambda_0 mr(Y_j)} - 1\right| \geq a\epsilon\right\}$$

$$\overset{(23)}{=} \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n+\lambda_0 mr(X_i)} + \frac{1}{m}\sum_{j=1}^{m}\frac{m}{n+\lambda_0 mr(Y_j)} - \int\frac{nf}{n+\lambda_0 mr}d\mu - \int\frac{mg}{n+\lambda_0 mr}d\mu\right| \geq a\epsilon\right\}$$

$$\leq \mathbb{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\frac{n}{n+\lambda_0 mr(X_i)} - \int\frac{nf}{n+\lambda_0 mr}d\mu\right| \geq \frac{a\epsilon}{2}\right\}$$

$$+ \mathbb{P}\left\{\left|\frac{1}{m}\sum_{j=1}^{m}\frac{m}{n+\lambda_0 mr(Y_j)} - \int\frac{mg}{n+\lambda_0 mr}d\mu\right| \geq \frac{a\epsilon}{2}\right\}$$

$$\leq 2\exp\left\{-\frac{na^2\epsilon^2}{2}\right\} + 2\exp\left\{-\frac{ma^2\epsilon^2}{2}\right\} \leq 2\exp\left\{-\frac{na^2\epsilon^2}{2}\right\} + 2\exp\left\{-\frac{na^2\epsilon^2}{2\tau}\right\} \leq 4\exp\left\{-\frac{na^2\epsilon^2}{2\tau}\right\},$$

$$(24)$$

where in the last line we applied Hoeffding's inequality for bounded random variables (see Wainwright, 2019, Equation 2.11) to $0 \leq \frac{m}{n+\lambda_0 mr(\cdot)} \leq \frac{n}{n+\lambda_0 mr(\cdot)} \leq 1$, and the assumption that $m \leq n \leq \tau m$. We can thus bound the moment of order $p$ of $|\widehat{\lambda} - \lambda_0|$ as

$$\mathbb{E}[|\widehat{\lambda} - \lambda_0|^p] = \int_0^\infty \mathbb{P}\{|\widehat{\lambda} - \lambda_0|^p \geq \epsilon\}d\epsilon = \int_0^\infty \mathbb{P}\{|\widehat{\lambda} - \lambda_0| \geq \epsilon^{\frac{1}{p}}\}d\epsilon$$

$$= \int_0^\infty \mathbb{P}\{|\widehat{\lambda} - \lambda_0| \geq \epsilon\}\, p\,\epsilon^{p-1}d\epsilon = \left(\frac{2\tau}{na^2}\right)^{\frac{p}{2}}\int_0^\infty \mathbb{P}\left\{|\widehat{\lambda} - \lambda_0| \geq \sqrt{\frac{2\tau}{na^2}}t\right\}\,p\,t^{p-1}dt$$

$$\leq 2p\left(\frac{2\tau}{na^2}\right)^{\frac{p}{2}}\int_0^\infty 2t^{p-1}e^{-t^2}dt = 2p\left(\frac{2\tau}{na^2}\right)^{\frac{p}{2}}\int_0^\infty t^{\frac{p}{2}-1}e^{-t}dt = 2p\left(\frac{2\tau}{na^2}\right)^{\frac{p}{2}}\Gamma\left(\frac{p}{2}\right),$$

where in the last equality we used the definition of the Gamma function. This completes the proof. $\square$

*Proof of Lemma 7.* We will start by proving $(i)$. The strategy is to relate the distribution of

$$\frac{S_n}{n} := \frac{1}{n}\sum_{i=1}^{n}\varphi(Z_{\sigma(i)}) - \int\varphi d\widehat{H}_{n,m},$$

where $\sigma$ is sampled from (3), to that of an analogous version that evolves over time according to an equivalent version of Algorithm 1. More precisely, writing $\sigma_t$ for the permutation at time $t \in \mathbb{N}$ of this new algorithm, we consider a procedure that at each time step $t$ samples $i \in [n]$ and $j \in \{n+1, \ldots, n+m\}$ uniformly at random, and switches $Z_{\sigma_t(i)}$ with $Z_{\sigma_t(j)}$ with probability

$$\tilde{p}_{i,j}^t := \mathbb{P}\{\text{switch } i \text{ and } j \text{ at time } t \mid i, j \text{ are selected}\} = \frac{\widehat{\lambda}nmr_i^t}{(n+\widehat{\lambda}mr_i^t)(n+\widehat{\lambda}mr_j^t)},$$

where $r_i^t := r(Z_{\sigma_t(i)})$ for all $i \in [n+m]$. As already outlined in Remark 1, this algorithm still targets the distribution (3) since $\tilde{p}_{i,j}^t / \tilde{p}_{j,i}^t = r_i^t / r_j^t$. Consider

$$
\begin{aligned}
\frac{S_n^t}{n} &:= \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma_t(i)}) - \int \varphi d\widehat{H}_{n,m} = \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma_t(i)}) - \sum_{i=1}^{n+m} \frac{\varphi(Z_i)}{n + \widehat{\lambda} m r(Z_i)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma_t(i)}) - \sum_{i=1}^{n+m} \frac{\varphi(Z_{\sigma_t(i)})}{n + \widehat{\lambda} m r(Z_{\sigma_t(i)})} \\
&= \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma_t(i)}) - \sum_{i=1}^{n} \frac{\varphi(Z_{\sigma_t(i)})}{n + \widehat{\lambda} m r(Z_{\sigma_t(i)})} - \sum_{j=n+1}^{n+m} \frac{\varphi(Z_{\sigma_t(j)})}{n + \widehat{\lambda} m r(Z_{\sigma_t(j)})} \\
&= \frac{1}{n} \left( \sum_{i=1}^{n} \frac{\widehat{\lambda} m r(Z_{\sigma_t(i)})}{n + \widehat{\lambda} m r(Z_{\sigma_t(i)})} \varphi(Z_{\sigma_t(i)}) - \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r(Z_{\sigma_t(j)})} \varphi(Z_{\sigma_t(j)}) \right) \\
&= \frac{1}{n} \left( \sum_{i=1}^{n} \frac{\widehat{\lambda} m r_i^t}{n + \widehat{\lambda} m r_i^t} \varphi_i^t - \sum_{j=n+1}^{n+m} \frac{n}{n + \widehat{\lambda} m r_j^t} \varphi_j^t \right),
\end{aligned}
\tag{25}
$$

where we further defined $\varphi_i^t = \varphi(Z_{\sigma_t(i)})$ for all $i \in [n+m]$. Supposing that the algorithm is initialised at the stationary distribution (3), we have $\mathbb{E}[(n^{-1} S_n)^2] = \mathbb{E}[(n^{-1} S_n^t)^2]$ for all $t \in \mathbb{N}$, therefore it is equivalent to prove the claim ($i$) for $S_n^t / n$. This approach offers the significant advantage of allowing us to leverage the Markov chain's dynamics to construct zero-mean random variables that can be linked to $S_n^t$ using the definition of $\tilde{p}_{i,j}^t$. In this regard, since switching the indices $i$ and $j$ at time $t$ gives $S_n^{t+1} - S_n^t = \varphi_j^t - \varphi_i^t$, we have

$$
\begin{aligned}
\mathbb{E}\left[ \left( \frac{S_n^{t+1}}{n} \right)^2 \mid \sigma_t, Z \right] &= \left\{ 1 - \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} \tilde{p}_{i,j}^t \right\} \left( \frac{S_n^t}{n} \right)^2 + \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} \left\{ \frac{S_n^t + \varphi_j^t - \varphi_i^t}{n} \right\}^2 \tilde{p}_{i,j}^t \\
&= \left( \frac{S_n^t}{n} \right)^2 + \frac{1}{n^3 m} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t)^2 \tilde{p}_{i,j}^t - \frac{2}{n^2 m} \frac{S_n^t}{n} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t) \tilde{p}_{i,j}^t.
\end{aligned}
$$

As a result, the law of total expectation and the fact that the procedure is initialised at stationarity imply that

$$
\begin{aligned}
0 &= \mathbb{E}\left[ \left( \frac{S_n^{t+1}}{n} \right)^2 \right] - \mathbb{E}\left[ \left( \frac{S_n^t}{n} \right)^2 \right] \\
&= \frac{1}{n^2} \mathbb{E}\left[ \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t)^2 \tilde{p}_{i,j}^t \right] - \frac{2}{n} \mathbb{E}\left[ \frac{S_n^t}{n} \left\{ \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t) \tilde{p}_{i,j}^t \right\} \right].
\end{aligned}
$$

Introduce the notation $q_i^t := \widehat{\lambda} m r_i^t / (n + \widehat{\lambda} m r_i^t)$ so that we may write $\tilde{p}_{i,j}^t = q_i^t (1 - q_j^t)$. With this definition it follows from (10) that we have $\sum_{i=1}^{n} q_i^t = \sum_{j=n+1}^{n+m} (1 - q_j^t)$. We therefore have

$$
\sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t) \tilde{p}_{i,j}^t = \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t) q_i^t (1 - q_j^t)
$$

$$= \left( \sum_{j=n+1}^{n+m} (1 - q_j^t) \right) \left( \sum_{i=1}^{n} \varphi_i^t q_i^t \right) - \left( \sum_{i=1}^{n} q_i^t \right) \left( \sum_{j=n+1}^{n+m} \varphi_j^t (1 - q_j^t) \right)$$

$$= \left( \sum_{i=1}^{n} q_i^t \right) \left( \sum_{i=1}^{n} \varphi_i^t q_i^t - \sum_{j=n+1}^{n+m} \varphi_j^t (1 - q_j^t) \right) = \left( \sum_{i=1}^{n} q_i^t \right) S_n^t.$$

Then, for $\|\varphi\|_\infty \le B < \infty$, it follows that

$$0 = \frac{1}{n^2} \mathbb{E} \left[ \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=n+1}^{n+m} (\varphi_i^t - \varphi_j^t)^2 \tilde{p}_{i,j}^t \right] - \frac{2}{m} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} q_i^t \right) \left( \frac{S_n^t}{n} \right)^2 \right]$$

$$\le \frac{4B^2}{n^2} - \frac{2}{m} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} q_i^t \right) \left( \frac{S_n^t}{n} \right)^2 \right].$$

Now observe that under the assumption that $0 < c \le r(x) \le C$ for all $x \in \mathcal{X}$, we have that $C^{-1} \le \widehat{\lambda} \le c^{-1}$, which implies

$$\frac{mc}{mc + nC} \le q_i \le \frac{mC}{mC + nc} \qquad \text{for all } i \in [n+m]. \tag{26}$$

It follows that

$$\frac{mc}{mc + nC} \mathbb{E} \left[ \left( \frac{S_n^t}{n} \right)^2 \right] \le \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} q_i^t \right) \left( \frac{S_n^t}{n} \right)^2 \right] \le \frac{2mB^2}{n^2},$$

which gives $\mathbb{E} \left[ \left( n^{-1} S_n^t \right)^2 \right] \le 2B^2(mc+nC)/cn^2$. Similar calculations show that we also have $\mathbb{E} \left[ \left( n^{-1} S_n^t \right)^2 \right] \le 2B^2(nc + mC)/cn^2$, and summing the two gives

$$\mathbb{E} \left[ \left( \frac{S_n^t}{n} \right)^2 \right] \le \frac{B^2(c+C)}{c} \frac{n+m}{n^2}.$$

This concludes the proof for part $(i)$, and further shows that $n^{-1} S_n^t \overset{\mathbb{P}}{\to} 0$ since $n/m \to \tau \in \mathbb{R}_+$.

We now move on to proving $(ii)$. Let $h$ be the density of the form $h = \frac{nf+mg}{n+\lambda_0 mr}$ for a suitable constant $\lambda_0 > 0$ such that $\int h \, d\mu = 1$. By the triangle inequality we have

$$\left| \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) - \int \varphi h_\infty \, d\mu \right| \le \left| \frac{1}{n} \sum_{i=1}^{n} \varphi(Z_{\sigma(i)}) - \int \varphi \, d\widehat{H}_{n,m} \right| + \left| \int \varphi \, d\widehat{H}_{n,m} - \int \varphi h \, d\mu \right| + \left| \int \varphi(h - h_\infty) \, d\mu \right|$$

$$=: (I) + (II) + (III),$$

and since $(I) \overset{\mathbb{P}}{\to} 0$ by part $(i)$, it remains to show that $(II)$ and $(III)$ likewise converge to zero in probability. We proceed by analysing each term individually. As for $(II)$, observe that

$$(II) = \left| \int \varphi \, d\widehat{H}_{n,m} - \int \varphi h \, d\mu \right| = \left| \sum_{i=1}^{n+m} \frac{\varphi(Z_i)}{n + \widehat{\lambda} mr(Z_i)} - \int \varphi \frac{nf + mg}{n + \lambda_0 mr} \, d\mu \right|$$

$$= \left| \sum_{i=1}^{n} \frac{\varphi(X_i)}{n + \widehat{\lambda}mr(X_i)} + \sum_{j=n+1}^{n+m} \frac{\varphi(Y_j)}{n + \widehat{\lambda}mr(Y_j)} - \int \frac{n\varphi}{n + \lambda_0 mr} f d\mu - \int \frac{m\varphi}{n + \lambda_0 mr} g d\mu \right|$$

$$\leq \left| \sum_{i=1}^{n} \frac{\varphi(X_i)}{n + \widehat{\lambda}mr(X_i)} - \int \frac{n\varphi}{n + \lambda_0 mr} f d\mu \right| + \left| \sum_{j=n+1}^{n+m} \frac{\varphi(Y_j)}{n + \widehat{\lambda}mr(Y_j)} - \int \frac{m\varphi}{n + \lambda_0 mr} g d\mu \right| =: (a) + (b).$$

Each of these terms is bounded using very similar arguments, so we restrict attention to $(b)$ here. We have

$$(b) = \left| \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \widehat{\lambda}mr(Y_j)} - \int \frac{m\varphi}{n + \lambda_0 mr} g d\mu \right|$$

$$\leq \left| \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \widehat{\lambda}mr(Y_j)} - \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \lambda_0 mr(Y_j)} \right| + \left| \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \lambda_0 mr(Y_j)} - \int \frac{m\varphi}{n + \lambda_0 mr} g d\mu \right|,$$

where the second term converges to zero in probability via Hoeffding's inequality, by combining a similar argument to that in (24) with the bound $\|\varphi\|_\infty \leq B$. As for the first term, we have

$$\left| \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \widehat{\lambda}mr(Y_j)} - \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{m\varphi(Y_j)}{n + \lambda_0 mr(Y_j)} \right|$$

$$= \left| (1 - \widehat{\lambda}/\lambda_0) \frac{1}{m} \sum_{j=n+1}^{n+m} \frac{\lambda_0 m^2 r(Y_j)\varphi(Y_j)}{\{n + \widehat{\lambda}mr(Y_j)\}\{n + \lambda_0 mr(Y_j)\}} \right|$$

$$\leq |\widehat{\lambda}/\lambda_0 - 1| \frac{1}{n} \sum_{j=n+1}^{n+m} |\varphi(Y_j)| \left( \frac{\lambda_0 mr(Y_j)}{n + \lambda_0 mr(Y_j)} \right) \left( \frac{n}{n + \widehat{\lambda}mr(Y_j)} \right) \leq \frac{Bm}{n} |\widehat{\lambda}/\lambda_0 - 1| \xrightarrow{\mathbb{P}} 0,$$

using Lemma 12, the fact that $C^{-1} \leq \lambda_0 \leq c^{-1}$ and $n/m \to \tau \in \mathbb{R}_+$. An almost identical argument applies also to $(a)$, which implies that $(II) \xrightarrow{\mathbb{P}} 0$. As for $(III)$, we can use again the fact that $\|\varphi\|_\infty \leq B$ to show that

$$(III) = \left| \int (h - h_\infty)\varphi d\mu \right| = \left| \int \left( \frac{\frac{n}{m}f + g}{\frac{n}{m} + \lambda_0 r} - \frac{\tau f + g}{\tau + \lambda_\infty r} \right) \varphi d\mu \right| \leq \|\varphi\|_\infty \int \left| \frac{\frac{n}{m}f + g}{\frac{n}{m} + \lambda_0 r} - \frac{\tau f + g}{\tau + \lambda_\infty r} \right| d\mu$$

$$\leq B \int \frac{|\tau - \frac{n}{m}|}{\frac{n}{m} + \lambda_0 r} f d\mu + B \int \frac{|\tau - \frac{n}{m}| + |\lambda_\infty - \lambda_0|r}{(\tau + \lambda_\infty r)(\frac{n}{m} + \lambda_0 r)} (\tau f + g) d\mu \to 0,$$

arguing as we did for (22). This concludes the proof. $\qquad\square$

*Proof of Proposition 8.* This proof borrows ideas from the proof of Lemma 4 in Gretton et al. (2012). Define the linear operator

$$T_{rf} : \begin{cases} \mathcal{H} \to \mathbb{R} \\ \varphi \mapsto \int \frac{\lambda_0 rf}{n/m + \lambda_0 r} \varphi \, d\mu. \end{cases}$$

Using the reproducing property of the RKHS, i.e. $\varphi(x) = \langle \varphi, k(\cdot, x) \rangle_\mathcal{H}$, we can show that this operator is

bounded, since for all $\varphi \in \mathcal{H}$ we have

$$|T_{rf}\varphi| \leq \int_{\mathcal{X}} \frac{\lambda_0 r(x) f(x)}{n/m + \lambda_0 r(x)} |\varphi(x)| d\mu(x) = \int_{\mathcal{X}} \frac{\lambda_0 r(x) f(x)}{n/m + \lambda_0 r(x)} |\langle \varphi, k(\cdot, x) \rangle_{\mathcal{H}}| d\mu(x)$$

$$\leq \|\varphi\|_{\mathcal{H}} \int_{\mathcal{X}} \sqrt{k(x,x)} \frac{\lambda_0 r(x) f(x)}{n/m + \lambda_0 r(x)} d\mu(x),$$

which shows that $|T_{rf}\varphi|/\|\varphi\|_{\mathcal{H}}$ is bounded uniformly in $\varphi$. The same is true for the linear operator

$$T_g : \begin{cases} \mathcal{H} \to \mathbb{R} \\ \varphi \mapsto \int \frac{g}{n/m + \lambda_0 r} \varphi \, d\mu, \end{cases}$$

hence the Riesz representation theorem implies that there exist $m_{rf}, m_g \in \mathcal{H}$ such that $T_{rf}\varphi = \langle m_{rf}, \varphi \rangle_{\mathcal{H}}$ and $T_g \varphi = \langle m_g, \varphi \rangle_{\mathcal{H}}$. Furthermore, using again the reproducing property of $\mathcal{H}$, we have that

$$m_{rf}(t) = \langle m_{rf}, k(t, \cdot) \rangle_{\mathcal{H}} = T_{rf} k(t, \cdot) = \int_{\mathcal{X}} \frac{\lambda_0 r(x) f(x)}{n/m + \lambda_0 r(x)} k(x,t) d\mu(x)$$

and, similarly,

$$m_g(t) = \int_{\mathcal{X}} \frac{g(x)}{n/m + \lambda_0 r(x)} k(x,t) d\mu(x).$$

This implies that

$$T_{\mathcal{F},r}^2(f,g) = \left( \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} \left| \int \frac{\lambda_0 r f - g}{n/m + \lambda_0 r} \varphi d\mu \right| \right)^2 = \left( \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} |T_{rf}\varphi - T_g \varphi| \right)^2$$

$$= \left( \sup_{\|\varphi\|_{\mathcal{H}} \leq 1} |\langle m_{rf} - m_g, \varphi \rangle_{\mathcal{H}}| \right)^2 = \|m_{rf} - m_g\|_{\mathcal{H}}^2 = \langle m_{rf}, m_{rf} \rangle_{\mathcal{H}} + \langle m_g, m_g \rangle_{\mathcal{H}} - 2\langle m_{rf}, m_g \rangle_{\mathcal{H}}$$

$$= \int_{\mathcal{X}} \frac{\lambda_0 r(t) f(t)}{n/m + \lambda_0 r(t)} \left( \int_{\mathcal{X}} \frac{\lambda_0 r(x) f(x)}{n/m + \lambda_0 r(x)} k(x,t) d\mu(x) \right) d\mu(t)$$

$$+ \int_{\mathcal{X}} \frac{g(t)}{n/m + \lambda_0 r(t)} \left( \int_{\mathcal{X}} \frac{g(x)}{n/m + \lambda_0 r(x)} k(x,t) d\mu(x) \right) d\mu(t)$$

$$- 2 \int_{\mathcal{X}} \frac{\lambda_0 r(t) f(t)}{n/m + \lambda_0 r(t)} \left( \int_{\mathcal{X}} \frac{g(x)}{n/m + \lambda_0 r(x)} k(x,t) d\mu(x) \right) d\mu(t)$$

$$= \mathbb{E}_{X,X'} \left[ \frac{\lambda_0^2 r(X) r(X')}{\{n/m + \lambda_0 r(X)\}\{n/m + \lambda_0 r(X')\}} k(X, X') \right]$$

$$+ \mathbb{E}_{Y,Y'} \left[ \frac{1}{\{n/m + \lambda_0 r(Y)\}\{n/m + \lambda_0 r(Y')\}} k(Y, Y') \right]$$

$$- 2 \mathbb{E}_{X,Y} \left[ \frac{\lambda_0 r(X)}{\{n/m + \lambda_0 r(X)\}\{n/m + \lambda_0 r(Y)\}} k(X, Y) \right],$$

where $X, X' \overset{\text{i.i.d}}{\sim} f$ independently of $Y, Y' \overset{\text{i.i.d}}{\sim} g$, and concludes the proof. $\qquad \square$

*Proof of Theorem 9.* We consider a kernel function $K : \mathbb{R} \to \mathbb{R}$ that satisfies the assumptions in Section 3,

namely $K \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) \cap L^4(\mathbb{R})$ and $K(0) = 1$. Based on this, we define the constants

$$\kappa_j(d) := \left( \int_{\mathbb{R}} |K(u)|^j \, du \right)^d < \infty \quad \text{for } j \in \{1, 2, 4\},$$

which are finite by assumption. Since the kernel is fixed in advance, we omit its dependence from the notation. We frequently use the analytical properties of the product kernel $k_\zeta$ with bandwidth $\zeta \geq 1$, which satisfies

$$\int_{\mathbb{R}^d} |k_\zeta(x, y)|^j \, dy = \zeta^{(j-1)d} \left( \int_{\mathbb{R}} |K(u)|^j \, du \right)^d = \kappa_j(d) \, \zeta^{(j-1)d} \quad \text{for all } x \in \mathbb{R}^d \text{ and } j \in \{1, 2, 4\}.$$

As a result, using the fact that $\max\{\|f\|_\infty, \|g\|_\infty\} \leq M$, it follows that for any $x \in \mathbb{R}^d$ we have

$$\mathbb{E}[|k_\zeta(x, X)|^j] = \int_{\mathbb{R}^d} |k_\zeta(x, y)|^j \, f(y) dy \leq M \int_{\mathbb{R}^d} |k_\zeta(x, y)|^j \, dy = M\kappa_j(d)\zeta^{(j-1)d} \quad \text{for } j \in \{1, 2, 4\},$$

and similarly for the expectation of $|k_\zeta(x, Y)|^j$. Furthermore, these bounds imply the same bounds on the quantities $\mathbb{E}[|k_\zeta(X_1, X_2)|^j]$, $\mathbb{E}[|k_\zeta(Y_1, Y_2)|^j]$ and $\mathbb{E}[|k_\zeta(X, Y)|^j]$ for all $j \in \{1, 2, 4\}$. Now, we already know that the DRPT controls the Type-I error at a nominal level $\alpha$, so we can bound the minimax separation $\rho_r^*$ by controlling its Type-II error. In order to do this, fix $\beta \in (0, 1 - \alpha)$, choose $H \geq 2\lceil \frac{1}{\alpha\beta} - 1 \rceil$, and suppose $(f, g) \in \mathcal{S}_\theta^r(\rho)$ satisfies

$$\text{MMD}^2_{r,k_\zeta}(f, g) \geq \max\left\{ 2|\mathbb{E}[U_{\text{id}} - U_\sigma] - \text{MMD}^2_{r,k_\zeta}(f, g)|, \left( \frac{8}{\alpha\beta} \text{Var}[U_\sigma - U_{\text{id}}] \right)^{1/2} \right\}, \tag{27}$$

where $U_\sigma = U(Z_\sigma)$ and $U_{\text{id}} = U(Z)$, with $U$ defined in (12) and $\sigma$ sampled from (3). Then, a double application of Markov's inequality shows that

$$\mathbb{P}\{p > \alpha\} = \mathbb{P}\left( 1 + \sum_{h=1}^H \mathbb{1}\{U_{\sigma^{(h)}} \geq U_{\text{id}}\} > (1 + H)\alpha \right) \leq \frac{1 + H\,\mathbb{P}\{U_\sigma \geq U_{\text{id}}\}}{(1 + H)\alpha}$$

$$\leq \frac{1}{(1 + H)\alpha} \left( 1 + \frac{H\,\text{Var}[U_\sigma - U_{\text{id}}]}{\{\mathbb{E}[U_\sigma - U_{\text{id}}]\}^2} \right) \leq \frac{1}{(1 + H)\alpha} \left( 1 + \frac{H\alpha\beta}{2} \right) \leq \beta.$$

Bounding the terms on the right-hand side of (27) is therefore sufficient to establish an upper bound on the minimax separation with respect to the squared shifted-MMD metric. However, since our goal is to characterise the separation in terms of the $L^2$ distance defined in (13), we must also relate $\text{MMD}^2_{r,k_\zeta}$ to this $L^2$ norm. The proof proceeds in two main steps: first, we analyse the expectation and variance terms appearing on the right-hand side of (27) and derive appropriate upper bounds. In the second step, we express the squared shifted-MMD as the sum of the square of the separation metric in (13) and the bias term $\|\psi_r - \varphi_\zeta * \psi_r\|_2^2$, which can be controlled using the smoothness assumptions associated with the Sobolev class. Combining these two steps yields an upper bound on $\rho_r^*$. Throughout the following we set $\zeta = n^{\frac{2}{4s+d}}$ so that in particular $n^{-2}\zeta^d = n^{-\frac{8s}{4s+d}} \leq 1$. Also, for parameters $a_1, \ldots, a_k$, we denote by $Q(a_1, \ldots, a_k)$ a constant that depends only on these parameters. Its value may change from line to line, but it may only

depend on on $a_1, \ldots, a_k$.

• **Mean and variance of $U_\sigma$**

We begin by analysing the second moment of the permuted U-statistic. Through the following, sums over $i$'s are to be intended for $i$ varying in $[n]$, sums over $j$'s are to be intended for $j$ varying in $\{n+1, \ldots, n+m\}$, and sums over $k$'s are to be intended for $k$ varying in $[n+m]$. As in the proof of Lemma 7, we introduce a stationary Markov chain whose stationary distribution coincides with the distribution of our permuted data. We do this by considering the equivalent version of Algorithm 1 which at each time step $t \in \mathbb{N}$ samples $i$ and $j$ uniformly at random, and switches $Z_{\sigma_t(i)}$ with $Z_{\sigma_t(j)}$ with probability

$$\tilde{p}_{i,j}^t := \mathbb{P}(\text{switch } i \text{ and } j \text{ at time } t \mid i, j \text{ are selected}) = \frac{\widehat{\lambda} n m r_i^t}{(n + \widehat{\lambda} m r_i^t)(n + \widehat{\lambda} m r_j^t)},$$

with $\widehat{\lambda}$ as in the statement of Lemma 7 and $r_k^t = r(Z_{\sigma_t(k)})$ for all $k \in [n+m]$. Write

$$K_{ij}^t := k_\zeta(Z_{\sigma_t(i)}, Z_{\sigma_t(j)}) \text{ and } K_i^t := k_\zeta(Z_{\sigma_t(i)}, \cdot),$$

$$q_i^t := \frac{\widehat{\lambda} m r_i^t}{n + \widehat{\lambda} m r_i^t} \text{ and } S_n^t := \sum_i q_i^t = \sum_j (1 - q_j^t),$$

where the last equality holds by (10). In particular we have $\tilde{p}_{i,j}^t = q_i^t(1 - q_j^t)$. We define the V-statistic

$$V_t = \frac{1}{n^2} \left\| \sum_i q_i^t K_i^t - \sum_j (1 - q_j^t) K_j^t \right\|_{\mathcal{H}}^2 =: \|G_t\|_{\mathcal{H}}^2 = \langle G_t, G_t \rangle_{\mathcal{H}},$$

which coincides with (11) evaluated on our data at time $t$ with kernel $k_\zeta$ and will be convenient for our analysis. We further define the U-statistic

$$U_t = V_t - \frac{\zeta^d}{n^2} \sum_i (q_i^t)^2 - \frac{\zeta^d}{n^2} \sum_j (1 - q_j^t)^2 = V_t - \frac{\zeta^d}{n^2} \sum_i (q_i^t)^2 - \frac{\zeta^d}{n^2} \sum_j \{1 + (q_j^t)^2 - 2q_j^t\}$$

$$= V_t - \frac{\zeta^d}{n^2} \sum_k (q_k^t)^2 + \frac{m\zeta^d}{n^2} - \frac{2\zeta^d}{n^2} \sum_j (1 - q_j^t) = V_t - \frac{2\zeta^d}{n^2} S_n^t \underbrace{- \frac{\zeta^d}{n^2} \sum_k (q_k^t)^2 + \frac{m\zeta^d}{n^2}}_{\text{permutation independent}},$$

which has the same distribution as $U_\sigma$ for each $t \in \mathbb{N}$. If we swap $i$ and $j$ at time $t$, then the difference $U_{t+1} - U_t$ is equal to

$$\|G_t + n^{-1}(K_j^t - K_i^t)\|_{\mathcal{H}}^2 - \|G_t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t) = \frac{2}{n}\langle G_t, K_j^t - K_i^t \rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t).$$

By stationarity, we therefore have

$$0 = nm \, \mathbb{E}[U_{t+1}^2 - U_t^2]$$

44

$$= \sum_{i,j} \mathbb{E}\left[q_i^t(1-q_j^t)\left(\left\{U_t + \frac{2}{n}\langle G_t, K_j^t - K_i^t\rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)\right\}^2 - U_t^2\right)\right]$$

$$= \sum_{i,j} \mathbb{E}\left[q_i^t(1-q_j^t) \cdot 2U_t\left\{\frac{2}{n}\langle G_t, K_j^t - K_i^t\rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)\right\}\right]$$

$$+ \sum_{i,j} \mathbb{E}\left[q_i^t(1-q_j^t)\left\{\frac{2}{n}\langle G_t, K_j^t - K_i^t\rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)\right\}^2\right], \tag{28}$$

which is the sum of a linear and a quadratic term. In order to simplify this right-hand side we now provide some useful identities. We notice that

$$\sum_{i,j} q_i^t(1-q_j^t)(K_j^t - K_i^t) = S_n^t\left\{\sum_j(1-q_j^t)K_j^t - \sum_i q_i^t K_i^t\right\} = -nS_n^t G_t,$$

$$\frac{1}{n^2}\sum_{i,j} q_i^t(1-q_j^t)\|K_j^t - K_i^t\|_{\mathcal{H}}^2 = \frac{1}{n^2}\sum_{i,j} q_i^t(1-q_j^t)(2\zeta^d - 2K_{ij}^t) = \frac{2\zeta^d}{n^2}(S_n^t)^2 - \frac{2}{n^2}\sum_{i,j} q_i^t(1-q_j^t)K_{ij}^t,$$

and

$$\frac{2\zeta^d}{n^2}\sum_{i,j} q_i^t(1-q_j^t)(q_j^t - q_i^t) = \frac{2\zeta^d}{n^2}S_n^t\left\{\sum_j q_j^t(1-q_j^t) - \sum_i(q_i^t)^2\right\} = \frac{2\zeta^d}{n^2}S_n^t\left\{m - S_n^t - \sum_k(q_k^t)^2\right\}.$$

Using these identities, the linear term in (28) gives

$$\sum_{i,j} \mathbb{E}\left[q_i^t(1-q_j^t) \cdot 2U_t\left\{\frac{2}{n}\langle G_t, K_j^t - K_i^t\rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)\right\}\right]$$

$$= \mathbb{E}\left[2U_t\left\{\frac{2}{n}\langle G_t, \sum_{i,j}(K_j^t - K_i^t)q_i^t(1-q_j^t)\rangle + \frac{1}{n^2}\sum_{i,j}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 q_i^t(1-q_j^t) - \frac{2\zeta^d}{n^2}\sum_{i,j}(q_j^t - q_i^t)q_i^t(1-q_j^t)\right\}\right]$$

$$= \mathbb{E}\left[2U_t\left\{-2S_n^t\left(U_t + \frac{2\zeta^d}{n^2}S_n^t + \frac{\zeta^d}{n^2}\sum_k(q_k^t)^2 - \frac{m\zeta^d}{n^2}\right)\right.\right.$$

$$\left.\left. + \frac{2\zeta^d}{n^2}(S_n^t)^2 - \frac{2}{n^2}\sum_{i,j} q_i^t(1-q_j^t)K_{ij}^t - \frac{2\zeta^d}{n^2}S_n^t\left(m - S_n^t - \sum_k(q_k^t)^2\right)\right\}\right]$$

$$= -4\mathbb{E}\left[S_n^t U_t^2\right] - 4\mathbb{E}\left[U_t\frac{1}{n^2}\sum_{i,j} q_i^t(1-q_j^t)K_{ij}^t\right]. \tag{29}$$

As for the quadratic term, it is useful to define $G_t^{-(i,j)} := G_t - n^{-1}\{q_i^t K_i^t - (1-q_j^t)K_j^t\}$ and write

$$\frac{2}{n}\langle G_t, K_j^t - K_i^t\rangle + \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)$$

$$= \frac{2}{n}\langle G_t - \frac{1}{n}\{q_i^t K_i^t - (1-q_j^t)K_j^t\}, K_j^t - K_i^t\rangle + \frac{2}{n^2}\langle q_i^t K_i^t - (1-q_j^t)K_j^t, K_j^t - K_i^t\rangle$$

$$+ \frac{1}{n^2}\|K_j^t - K_i^t\|_{\mathcal{H}}^2 - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)$$

$$= \frac{2}{n}\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle + \frac{2}{n^2}\langle q_i^t K_i^t - (1-q_j^t)K_j^t, K_j^t - K_i^t\rangle + \frac{2}{n^2}(\zeta^d - K_{ij}^t) - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)$$

$$= \frac{2}{n}\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle + \frac{2}{n^2}(q_i^t K_{ij}^t - \zeta^d q_i^t - \zeta^d(1-q_j^t) + (1-q_j^t)K_{ij}^t) + \frac{2}{n^2}(\zeta^d - K_{ij}^t) - \frac{2\zeta^d}{n^2}(q_j^t - q_i^t)$$

$$= \frac{2}{n}\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle + \frac{2}{n^2}(q_i^t - q_j^t)K_{ij}^t. \tag{30}$$

Combining (28), (29), (30) gives

$$4\mathbb{E}\left[S_n^t U_t^2\right] = -4\mathbb{E}\left[U_t \frac{1}{n^2}\sum_{i,j} q_i^t(1-q_j^t)K_{ij}^t\right] + \sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\left\{\frac{2}{n}\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle + \frac{2}{n^2}(q_i^t - q_j^t)K_{ij}^t\right\}^2\right]$$

$$\leq 4\mathbb{E}\left[|U_t|\frac{1}{n^2}\sum_{i,j} q_i^t(1-q_j^t)|K_{ij}^t|\right] + \sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\left\{\frac{2}{n}\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle + \frac{2}{n^2}(q_i^t - q_j^t)K_{ij}^t\right\}^2\right]$$

$$\leq 4\sqrt{\mathbb{E}[U_t^2]\mathbb{E}\left[n^{-4}\left\{\sum_{i,j} q_i^t(1-q_j^t)|K_{ij}^t|\right\}^2\right]} + \frac{8}{n^2}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle^2\right]$$

$$+ \frac{8}{n^4}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)(q_i^t - q_j^t)^2(K_{ij}^t)^2\right]$$

$$\leq 4\sqrt{\mathbb{E}[U_t^2]\mathbb{E}[n^{-2}\sum_{i,j}\{q_i^t(1-q_j^t)K_{ij}^t\}^2]} + \frac{8}{n^2}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle^2\right]$$

$$+ \frac{8}{n^4}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)(q_i^t - q_j^t)^2(K_{ij}^t)^2\right]$$

$$\leq 12\max\left\{\sqrt{\mathbb{E}[U_t^2]\mathbb{E}[n^{-2}\sum_{i,j}\{q_i^t(1-q_j^t)K_{ij}^t\}^2]}, \frac{2}{n^2}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle^2\right],\right.$$

$$\left.\frac{2}{n^4}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)(q_i^t - q_j^t)^2(K_{ij}^t)^2\right]\right\}. \tag{31}$$

We will now bound each of the three quantities inside the maximum separately. If the first quantity reaches the maximum, we can use $m \leq n \leq \tau m$, $q_k^t \in [0,1]$ and $S_n^t \geq \frac{nmc}{nC+mc}$ (see (26) in the proof of Lemma 7) to show that

$$\frac{nmc}{nC+mc}\mathbb{E}\left[U_t^2\right] \leq \mathbb{E}\left[S_n^t U_t^2\right] \leq 3\sqrt{\mathbb{E}[U_t^2]\mathbb{E}[n^{-2}\sum_{i,j}\{q_i^t(1-q_j^t)K_{ij}^t\}^2]}$$

$$\leq 3\sqrt{\mathbb{E}[U_t^2]\mathbb{E}[n^{-2}\sum_{i,j}(K_{ij}^t)^2]} \leq 3\sqrt{\mathbb{E}[U_t^2]\mathbb{E}[n^{-2}\sum_{k_1 \neq k_2}(K_{k_1,k_2}^t)^2]}$$

$$= 3\sqrt{\mathbb{E}[U_t^2]\mathbb{E}\left[n^{-2}\left\{\sum_{i_1\neq i_2}k_\zeta^2(X_{i_1},X_{i_2}) + \sum_{j_1\neq j_2}k_\zeta^2(Y_{j_1},Y_{j_2}) + 2\sum_{i,j}k_\zeta^2(X_i,Y_j)\right\}\right]}$$

$$= 3\sqrt{\mathbb{E}[U_t^2]\mathbb{E}\left[n^{-2}\left\{n(n-1)k_\zeta^2(X_1,X_2) + m(m-1)k_\zeta^2(Y_1,Y_2) + 2nmk_\zeta^2(X_1,Y_1)\right\}\right]}$$

$$\leq 3\sqrt{\mathbb{E}[U_t^2]\left\{\mathbb{E}[k_\zeta^2(X_1,X_2)] + \mathbb{E}[k_\zeta^2(Y_1,Y_2)] + 2\mathbb{E}[k_\zeta^2(X_1,Y_1)]\right\}} \leq 6\sqrt{\mathbb{E}[U_t^2]\,M\kappa_2(d)\zeta^d}. \quad (32)$$

Note that in the fourth inequality we bound $\sum_{i,j}(K_{i,j}^t)^2$ by $\sum_{k_1\neq k_2}(K_{k_1,k_2}^t)^2$, the latter being more advantageous to analyse due to its permutation invariance. Making (32) explicit for the expectation of $U_t^2$ shows that there exists a constant $Q_0 \equiv Q_0(c,C,d,\tau,M)$ such that $\mathbb{E}[U_t^2] \leq Q_0\frac{\zeta^d}{n^2}$, in the case where the first quantity in the maximum in (31) dominates. As for the case when the third quantity dominates, we can argue exactly as in (32) to show that

$$n^{-4}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)(q_i^t-q_j^t)^2(K_{ij}^t)^2\right] \leq \frac{4M\kappa_2(d)\zeta^d}{n^2}, \quad (33)$$

which implies that $\mathbb{E}[U_t^2] \leq Q_0\frac{\zeta^d}{n^3}$. Finally, as for the term involving $G_t^{-(i,j)}$, we have

$$n^{-2}\sum_{i,j}\mathbb{E}\left[q_i^t(1-q_j^t)\langle G_t^{-(i,j)}, K_j^t - K_i^t\rangle^2\right] \leq 2n^{-2}\sum_{i,j}\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle^2\right] + 2n^{-2}\sum_{i,j}\mathbb{E}\left[\langle G_t^{-(i,j)}, K_j^t\rangle^2\right],$$

and we can analyse the evolution of the right hand side through a version of Algorithm 1 where indices $i$ and $j$ remain fixed. In other words, the Algorithm works the same, but we are just allowed to swap indices $\tilde{i} \neq i$ with $\tilde{j} \neq j$ with probability $q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)$. We show in Lemma 13 that this procedure still preserves the stationary distribution. The two terms above can be bounded by almost identical arguments so we will restrict attention to the first. In analysing the evolution of $\langle G_t^{-(i,j)}, K_i^t\rangle^2$, it is useful to write

$$\sum_{\tilde{i}\neq i}\sum_{\tilde{j}\neq j}q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)(K_{\tilde{j}}^t - K_{\tilde{i}}^t) = \left(\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t\right)\left(\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t\right) - \left(\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)\right)\left(\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t\right)$$

$$= -nS_n^t G_t^{-(i,j)} + (1-q_j^t)\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t - q_i^t\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t.$$

Under stationarity, we thus have

$$0 = nm\,\mathbb{E}[\langle G_{t+1}^{-(i,j)}, K_i^t\rangle^2 - \langle G_t^{-(i,j)}, K_i^t\rangle^2]$$

$$= \sum_{\tilde{i}\neq i}\sum_{\tilde{j}\neq j}\mathbb{E}\left[q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)\left\{\langle G_t^{-(i,j)} + n^{-1}(K_{\tilde{j}}^t - K_{\tilde{i}}^t), K_i^t\rangle^2 - \langle G_t^{-(i,j)}, K_i^t\rangle^2\right\}\right]$$

$$= \sum_{\tilde{i}\neq i}\sum_{\tilde{j}\neq j}\mathbb{E}\left[n^{-2}q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)\langle K_{\tilde{j}}^t - K_{\tilde{i}}^t, K_i^t\rangle^2\right] + 2n^{-1}\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle\langle\sum_{\tilde{i}\neq i}\sum_{\tilde{j}\neq j}q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)(K_{\tilde{j}}^t - K_{\tilde{i}}^t), K_i^t\rangle\right]$$

$$= \sum_{\tilde{i}\neq i}\sum_{\tilde{j}\neq j}\mathbb{E}\left[n^{-2}q_{\tilde{i}}^t(1-q_{\tilde{j}}^t)\langle K_{\tilde{j}}^t - K_{\tilde{i}}^t, K_i^t\rangle^2\right] - 2\mathbb{E}[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2]$$

47

$$+ 2n^{-1}\,\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle\langle(1-q_j^t)\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t - q_i^t\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t, K_i^t\rangle\right]$$

$$\leq \frac{2(n-1)}{n^2}\sum_{\tilde{j}\neq j}\mathbb{E}\left[\langle K_{\tilde{j}}^t, K_i^t\rangle^2\right] + \frac{2(m-1)}{n^2}\sum_{\tilde{i}\neq i}\mathbb{E}\left[\langle K_{\tilde{i}}^t, K_i^t\rangle^2\right] - 2\mathbb{E}[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2]$$

$$+ 2n^{-1}\,\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle\langle(1-q_j^t)\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t - q_i^t\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t, K_i^t\rangle\right]$$

$$\leq 4\max_{k_1\neq k_2}\mathbb{E}[k_\zeta^2(Z_{k_1}, Z_{k_2})] - 2\mathbb{E}[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2]$$

$$+ 2n^{-1}\,\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle\langle(1-q_j^t)\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t - q_i^t\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t, K_i^t\rangle\right]$$

$$\leq 4M\kappa_2(d)\zeta^d - 2\mathbb{E}\left[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]$$

$$+ 2n^{-1}\sqrt{\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]\mathbb{E}\left[\langle(1-q_j^t)\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t - q_i^t\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t, K_i^t\rangle^2\right]}$$

$$\leq 4M\kappa_2(d)\zeta^d - 2\mathbb{E}\left[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]$$

$$+ 2\sqrt{2}\sqrt{\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]\mathbb{E}\left[\langle n^{-1}\sum_{\tilde{i}\neq i}q_{\tilde{i}}^t K_{\tilde{i}}^t, K_i^t\rangle^2 + \langle n^{-1}\sum_{\tilde{j}\neq j}(1-q_{\tilde{j}}^t)K_{\tilde{j}}^t, K_i^t\rangle^2\right]}$$

$$\leq 4M\kappa_2(d)\zeta^d - 2\mathbb{E}\left[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2\right] + 4\sqrt{\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]\max_{k_1\neq k_2}\mathbb{E}[k_\zeta^2(Z_{k_1}, Z_{k_2})]}$$

$$\leq 4M\kappa_2(d)\zeta^d - 2\mathbb{E}\left[S_n^t\langle G_t^{-(i,j)}, K_i^t\rangle^2\right] + 4\sqrt{M\kappa_2(d)\zeta^d\,\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i^t\rangle^2\right]}.$$

Using again the fact that $S_n^t \geq \frac{nmc}{nC+mc}$, we can employ the previous calculations to show that

$$\frac{2nmc}{nC+mc}\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i\rangle^2\right] \leq 2\mathbb{E}\left[S_n^t\langle G_t^{-(i,j)}, K_i\rangle^2\right] \leq 4M\kappa_2(d)\zeta^d + 4\sqrt{M\kappa_2(d)\zeta^d\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i\rangle^2\right]}$$

$$\leq 8\max\left\{M\kappa_2(d)\zeta^d, \sqrt{M\kappa_2(d)\zeta^d\mathbb{E}\left[\langle G_t^{-(i,j)}, K_i\rangle^2\right]}\right\}.$$

Thus, either $\mathbb{E}[\langle G_t^{-(i,j)}, K_i\rangle^2] \leq \frac{4(C+c)\tau M\kappa_2(d)\zeta^d}{cn}$ or $\frac{2nmc}{nC+mc}\mathbb{E}[\langle G_t^{-(i,j)}, K_i\rangle^2] \leq 8\sqrt{M\kappa_2(d)\zeta^d\mathbb{E}[\langle G_t^{-(i,j)}, K_i\rangle^2]}$, which implies that $\mathbb{E}[\langle G_t^{-(i,j)}, K_i\rangle^2] \leq \frac{16(C+c)^2\tau^2 M\kappa_2(d)\zeta^d}{c^2 n^2}$. This is sufficient to show that $\mathbb{E}[U_t^2] \leq Q_0\frac{\zeta^d}{n^2}$ also when the second quantity in (31) attains the maximum. Combining this with (32) and (33) gives

$$\max\left\{\mathbb{E}^2[|U_\sigma|], \mathrm{Var}[U_\sigma]\right\} \leq \mathbb{E}[U_\sigma^2] \leq Q_0\frac{\zeta^d}{n^2}, \tag{34}$$

and concludes the analysis for the moments for the permuted U-statistic.

- **Mean and variance of $U_{\mathrm{id}}$**

We now proceed with the analysis of the moments of $U_{\mathrm{id}}$. This would be straightforward if we used the classical normalisation factors $\frac{1}{n(n-1)}$ and $\frac{1}{m(m-1)}$ instead of $\frac{1}{n^2}$ and $\frac{1}{m^2}$, and if $\widehat{\lambda}$ were not random. The former problem is easy to address. Define

$$
\begin{aligned}
\tilde{U}_{\mathrm{id}} = {} & \frac{1}{n(n-1)} \sum_{i_1 \neq i_2 = 1}^{n} \frac{\widehat{\lambda}^2 r(X_{i_1}) r(X_{i_2}) k_\zeta(X_{i_1}, X_{i_2})}{\{\frac{n}{m} + \widehat{\lambda} r(X_{i_1})\}\{\frac{n}{m} + \widehat{\lambda} r(X_{i_2})\}} \\
& + \frac{1}{m(m-1)} \sum_{j_1 \neq j_2 = 1}^{m} \frac{k_\zeta(Y_{j_1}, Y_{j_2})}{\{\frac{n}{m} + \widehat{\lambda} r(Y_{j_1})\}\{\frac{n}{m} + \widehat{\lambda} r(Y_{j_2})\}} - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\widehat{\lambda} r(X_i) k_\zeta(X_i, Y_j)}{\{\frac{n}{m} + \widehat{\lambda} r(X_i)\}\{\frac{n}{m} + \widehat{\lambda} r(Y_j)\}},
\end{aligned} \tag{35}
$$

and observe that $\tilde{U}_{\mathrm{id}} - U_{\mathrm{id}}$ equals

$$
\frac{1}{n^2(n-1)} \sum_{i_1 \neq i_2 = 1}^{n} \frac{\widehat{\lambda}^2 r(X_{i_1}) r(X_{i_2}) k_\zeta(X_{i_1}, X_{i_2})}{\{\frac{n}{m} + \widehat{\lambda} r(X_{i_1})\}\{\frac{n}{m} + \widehat{\lambda} r(X_{i_2})\}} + \frac{1}{m^2(m-1)} \sum_{j_1 \neq j_2 = 1}^{m} \frac{k_\zeta(Y_{j_1}, Y_{j_2})}{\{\frac{n}{m} + \widehat{\lambda} r(Y_{j_1})\}\{\frac{n}{m} + \widehat{\lambda} r(Y_{j_2})\}}.
$$

Due to the boundedness assumption on $r$, this implies that there exists a constant $Q_1 \equiv Q_1(c, C, \tau) > 0$ such that

$$
\begin{aligned}
|\mathbb{E}[U_{\mathrm{id}}] - \mathrm{MMD}^2_{r,k_\zeta}(f,g)| &\leq |\mathbb{E}[\tilde{U}_{\mathrm{id}}] - \mathrm{MMD}^2_{r,k_\zeta}(f,g)| + \mathbb{E}[|U_{\mathrm{id}} - \tilde{U}_{\mathrm{id}}|] \\
&\leq |\mathbb{E}[\tilde{U}_{\mathrm{id}}] - \mathrm{MMD}^2_{r,k_\zeta}(f,g)| + \mathbb{E}\left[ \frac{Q_1}{n^2(n-1)} \sum_{i_1 \neq i_2 = 1}^{n} |k_\zeta(X_{i_1}, X_{i_2})| + \frac{Q_1}{m^2(m-1)} \sum_{j_1 \neq j_2 = 1}^{m} |k_\zeta(Y_{j_1}, Y_{j_2})| \right] \\
&\leq |\mathbb{E}[\tilde{U}_{\mathrm{id}}] - \mathrm{MMD}^2_{r,k_\zeta}(f,g)| + \frac{Q_1 M \kappa_1(d)}{n},
\end{aligned} \tag{36}
$$

and

$$
\begin{aligned}
\mathrm{Var}[U_{\mathrm{id}}] &\leq 2\,\mathrm{Var}[\tilde{U}_{\mathrm{id}}] + 2\,\mathrm{Var}[U_{\mathrm{id}} - \tilde{U}_{\mathrm{id}}] \\
&\leq 2\,\mathrm{Var}[\tilde{U}_{\mathrm{id}}] + \frac{Q_1}{n^2} \left( \max_{i_1 \neq i_2} \mathbb{E}[k_\zeta^2(X_{i_1}, X_{i_2})] + \max_{j_1 \neq j_2} \mathbb{E}[k_\zeta^2(Y_{j_1}, Y_{j_2})] \right) \leq 2\,\mathrm{Var}[\tilde{U}_{\mathrm{id}}] + \frac{Q_1 M \kappa_2(d) \zeta^d}{n^2}.
\end{aligned} \tag{37}
$$

Thus, since the second terms on the right-hand side of (36) and (37) are smaller or equal in order to that in (34), controlling the mean and variance of $\tilde{U}_{\mathrm{id}}$ is sufficient to bound those of $U_{\mathrm{id}}$.

- **Mean of $\tilde{U}_{\mathrm{id}}$**

We now address the harder problem of having $\widehat{\lambda}$ in $\tilde{U}_{\mathrm{id}}$, instead of the non-random quantity $\lambda_0$ appearing in the definition of $\mathrm{MMD}^2_{r,k_\zeta}(f,g)$. To overcome this issue, for $\lambda > 0$ define

$$
\begin{aligned}
G(\lambda) := {} & \frac{1}{n(n-1)} \sum_{i_1 \neq i_2 = 1}^{n} \frac{\lambda^2 r(X_{i_1}) r(X_{i_2}) k_\zeta(X_{i_1}, X_{i_2})}{\{n/m + \lambda r(X_{i_1})\}\{n/m + \lambda r(X_{i_2})\}} \\
& + \frac{1}{m(m-1)} \sum_{j_1 \neq j_2 = 1}^{m} \frac{k_\zeta(Y_{j_1}, Y_{j_2})}{\{n/m + \lambda r(Y_{j_1})\}\{n/m + \lambda r(Y_{j_2})\}} - \frac{2}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\lambda r(X_i) k_\zeta(X_i, Y_j)}{\{n/m + \lambda r(X_i)\}\{n/m + \lambda r(Y_j)\}},
\end{aligned} \tag{38}
$$

so that $\tilde{U}_{\mathrm{id}} = G(\widehat{\lambda})$. By expanding $G$ around $\lambda_0$ using a Taylor sum up to the second order we get $G(\widehat{\lambda}) = G(\lambda_0) + (\widehat{\lambda} - \lambda_0)\, G'(\lambda_0) + \frac{1}{2}(\widehat{\lambda} - \lambda_0)^2\, G''(\check{\lambda})$, where the random $\check{\lambda}$ is in between $\widehat{\lambda}$ and $\lambda_0$. We can use this identity to bound the mean of $\tilde{U}_{\mathrm{id}}$ as follows:

$$|\mathbb{E}[G(\widehat{\lambda})] - \mathrm{MMD}^2_{r,k_\zeta}(f,g)| = |\mathbb{E}[G(\lambda_0)] - \mathrm{MMD}^2_{r,k_\zeta}(f,g) + \mathbb{E}[(\widehat{\lambda} - \lambda_0)G'(\lambda_0)] + \frac{1}{2}\mathbb{E}[(\widehat{\lambda} - \lambda_0)^2 G''(\check{\lambda})]|$$

$$\leq |\mathbb{E}[(\widehat{\lambda} - \lambda_0)G'(\lambda_0)]| + |\frac{1}{2}\mathbb{E}[(\widehat{\lambda} - \lambda_0)^2 G''(\check{\lambda})]| \leq \sqrt{\mathbb{E}[(\widehat{\lambda} - \lambda_0)^2]\mathbb{E}[\{G'(\lambda_0)\}^2]} + \frac{1}{2}\sqrt{\mathbb{E}[(\widehat{\lambda} - \lambda_0)^4]\mathbb{E}[\{G''(\check{\lambda})\}^2]}$$

$$\leq Q_1\sqrt{\frac{1}{n}\mathbb{E}[\{G'(\lambda_0)\}^2]} + Q_1\sqrt{\frac{1}{n^2}\mathbb{E}[\{G''(\check{\lambda})\}^2]}$$

$$= Q_1\sqrt{\frac{1}{n}\{\mathbb{E}[G'(\lambda_0)]\}^2 + \frac{1}{n}\operatorname{Var}[G'(\lambda_0)]} + Q_1\sqrt{\frac{1}{n^2}\mathbb{E}[\{G''(\check{\lambda})\}^2]}. \tag{39}$$

Note that in the first bound we used the triangle inequality together with the fact that the U-statistic $G(\lambda_0)$ is unbiased for $\mathrm{MMD}^2_{r,k_\zeta}$, while in the third one we used Lemma 12. Now, as for the last term, we can use the fact that $0 < c \leq r(\cdot) \leq C$ to show that the first and second derivatives of $\lambda \mapsto \frac{\lambda^j}{\{n/m + \lambda r(x)\}\{n/m + \lambda r(y)\}}$ are uniformly bounded for all $\lambda > 0$, $x, y \in \mathbb{R}^d$ and $j \in \{0, 1, 2\}$. We can thus argue as we did for the second term in (37) to get

$$\mathbb{E}[\{G''(\check{\lambda})\}^2] \leq Q_1 \max_{k_1 \neq k_2} \mathbb{E}\left[k_\zeta^2(Z_{k_1}, Z_{k_2})\right] \leq Q_1 M \kappa_2(d)\zeta^d. \tag{40}$$

As for the first two terms, observe that $G'(\lambda_0)$ is a two-sample second-order U-statistic with defining kernel $\check{h}(x_1, x_2, y_1, y_2) = b_{XX}(x_1, x_2)k_\zeta(x_1, x_2) + b_{YY}(y_1, y_2)k_\zeta(y_1, y_2) + b_{XY}(x_1, y_2)k_\zeta(x_1, y_2) + b_{XY}(x_2, y_1)k_\zeta(x_2, y_1)$, where

$$\begin{cases} b_{XX}(x_1, x_2) := \frac{2\lambda_0 r(x_1)r(x_2)}{(n/m + \lambda_0 r(x_1))(n/m + \lambda_0 r(x_2))} - \frac{\lambda_0^2 r^2(x_1)r(x_2)}{(n/m + \lambda_0 r(x_1))^2(n/m + \lambda_0 r(x_2))} - \frac{\lambda_0^2 r(x_1)r^2(x_2)}{(n/m + \lambda_0 r(x_1))(n/m + \lambda_0 r(x_2))^2} \\ b_{YY}(y_1, y_2) := -\frac{r(y_1)}{(n/m + \lambda_0 r(y_1))^2(n/m + \lambda_0 r(y_2))} - \frac{r(y_2)}{(n/m + \lambda_0 r(y_1))(n/m + \lambda_0 r(y_2))^2} \\ b_{XY}(x_1, y_2) := -\frac{r(x_1)}{(n/m + \lambda_0 r(x_1))(n/m + \lambda_0 r(y_2))} + \frac{\lambda_0 r^2(x_1)}{(n/m + \lambda_0 r(x_1))^2(n/m + \lambda_0 r(y_2))} + \frac{\lambda_0 r(x_1)r(y_2)}{(n/m + \lambda_0 r(x_1))(n/m + \lambda_0 r(y_2))^2}. \end{cases}$$

Using again the boundedness assumption on $r(\cdot)$ we have that $\max\{|b_{XX}(\cdot, \cdot)|, |b_{YY}(\cdot, \cdot)|, |b_{XY}(\cdot, \cdot)|\} \leq Q_1$. Now, arguing as in Kim et al. (2022, Equation (69)), we can use Lee (1990, Equation 2 pag. 38) to show that

$$\operatorname{Var}[G'(\lambda_0)] \leq Q_2 \left\{ \frac{\check{\sigma}_{10}^2}{n} + \frac{\check{\sigma}_{01}^2}{n} + \left(\frac{1}{n} + \frac{1}{m}\right)^2 \check{\sigma}_{22}^2 \right\}$$

for a sufficiently large universal constant $Q_2 > 0$, where $\check{\sigma}_{10}^2 = \operatorname{Var}_{X_1}[\mathbb{E}_{X_2, Y_1, Y_2}\{\check{h}(X_1, X_2, Y_1, Y_2)\}]$, $\check{\sigma}_{01}^2 = \operatorname{Var}_{Y_1}[\mathbb{E}_{X_1, X_2, Y_2}\{\check{h}(X_1, X_2, Y_1, Y_2)\}]$, $\check{\sigma}_{22}^2 = \operatorname{Var}_{X_1, X_2, Y_1, Y_2}[\check{h}(X_1, X_2, Y_1, Y_2)]$. It is immediate to show that $\check{\sigma}_{22}^2 \leq Q_1 M \kappa_2(d)\zeta^d$ arguing as in (40), while

$$\check{\sigma}_{10}^2 = \operatorname{Var}_{X_1}[\mathbb{E}_{X_2}\{b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)\} + \mathbb{E}_{Y_2}\{b_{XY}(X_1, Y_2)k_\zeta(X_1, Y_2)\}]$$

$$\leq 2\mathbb{E}_{X_1}[\mathbb{E}_{X_2}^2\{b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)\} + \mathbb{E}_{Y_2}^2\{b_{XY}(X_1, Y_2)k_\zeta(X_1, Y_2)\}]$$

$$\leq 2\mathbb{E}_{X_1, X_2}[b_{XX}^2(X_1, X_2)k_\zeta^2(X_1, X_2)] + 2\mathbb{E}_{X_1, Y_2}[b_{XY}^2(X_1, Y_2)k_\zeta^2(X_1, Y_2)] \leq Q_1 M \kappa_2(d)\zeta^d.$$

The same holds true for $\check{\sigma}_{01}^2$, and shows that $n^{-1}\operatorname{Var}[G'(\lambda_0)] \leq Q_1 M \kappa_2(d)n^{-2}\zeta^d$. Finally, we bound

the expectation of $G'(\lambda_0)$ by $\|\varphi_\zeta * \psi_r\|_2$ up to constants, where $*$ stands for the convolution operator, $\psi_r = \frac{\lambda_0 mrf - mg}{n + \lambda_0 mr}$ and $\varphi_\zeta(x - y) = k_\zeta(x, y)$. In this regard, we have

$$\mathbb{E}[G'(\lambda_0)] = \mathbb{E}[b_{XX}(X_1, X_2)] + \mathbb{E}[b_{YY}(Y_1, Y_2)] + 2\mathbb{E}[b_{XX}(X, Y)]$$

$$= -\mathbb{E}\left[\frac{2\lambda_0^2 r^2(X_1) r(X_2) \varphi_\zeta(X_1 - X_2)}{(\frac{n}{m} + \lambda_0 r(X_1))^2 (\frac{n}{m} + \lambda_0 r(X_2))}\right] + \mathbb{E}\left[\frac{2\lambda_0 r(X_1) r(X_2) \varphi_\zeta(X_1 - X_2)}{(\frac{n}{m} + \lambda_0 r(X_1))(\frac{n}{m} + \lambda_0 r(X_2))}\right]$$

$$- \mathbb{E}\left[\frac{2r(Y_1) \varphi_\zeta(Y_1 - Y_2)}{(\frac{n}{m} + \lambda_0 r(Y_1))^2 (\frac{n}{m} + \lambda_0 r(Y_2))}\right] - \mathbb{E}\left[\frac{2r(X) \varphi_\zeta(X - Y)}{(\frac{n}{m} + \lambda_0 r(X))(\frac{n}{m} + \lambda_0 r(Y))}\right]$$

$$+ \mathbb{E}\left[\frac{2\lambda_0 r^2(X) \varphi_\zeta(X - Y)}{(\frac{n}{m} + \lambda_0 r(X))^2 (\frac{n}{m} + \lambda_0 r(Y))}\right] + \mathbb{E}\left[\frac{2\lambda_0 r(X) r(Y) \varphi_\zeta(X - Y)}{(\frac{n}{m} + \lambda_0 r(X))(\frac{n}{m} + \lambda_0 r(Y))^2}\right]$$

$$= \int_{\mathbb{R}^d} \frac{2r(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))} \left(\int_{\mathbb{R}^d} \varphi_\zeta(x - y) \left\{\frac{\lambda_0 r(y) f(y)}{(\frac{n}{m} + \lambda_0 r(y))} - \frac{g(y)}{(\frac{n}{m} + \lambda_0 r(y))}\right\} dy\right) dx$$

$$- \int_{\mathbb{R}^d} \frac{2\lambda_0 r^2(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))^2} \left(\int_{\mathbb{R}^d} \varphi_\zeta(x - y) \left\{\frac{\lambda_0 r(y) f(y)}{(\frac{n}{m} + \lambda_0 r(y))} - \frac{g(y)}{(\frac{n}{m} + \lambda_0 r(y))}\right\} dy\right) dx$$

$$+ \int_{\mathbb{R}^d} \frac{2r(x) g(x)}{(\frac{n}{m} + \lambda_0 r(x))^2} \left(\int_{\mathbb{R}^d} \varphi_\zeta(x - y) \left\{\frac{\lambda_0 r(y) f(y)}{(\frac{n}{m} + \lambda_0 r(y))} - \frac{g(y)}{(\frac{n}{m} + \lambda_0 r(y))}\right\} dy\right) dx$$

$$= \int_{\mathbb{R}^d} \left\{\frac{2r(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))} - \frac{2\lambda_0 r^2(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))^2} + \frac{2r(x) g(x)}{(\frac{n}{m} + \lambda_0 r(x))^2}\right\} (\varphi_\zeta * \psi_r)(x) dx$$

$$\leq \int_{\mathbb{R}^d} \left|\frac{2r(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))} - \frac{2\lambda_0 r^2(x) f(x)}{(\frac{n}{m} + \lambda_0 r(x))^2} + \frac{2r(x) g(x)}{(\frac{n}{m} + \lambda_0 r(x))^2}\right| |(\varphi_\zeta * \psi_r)(x)| dx \leq Q_1 \|\varphi_\zeta * \psi_r\|_2. \quad (41)$$

Combining this with (36), (39) and (40) enables to conclude the analysis of the first moment of $U_{\mathrm{id}}$, showing that

$$|\mathbb{E}[U_{\mathrm{id}}] - \mathrm{MMD}_{r,k_\zeta}^2(f, g)| \leq Q_0 \sqrt{n^{-2} \zeta^d + n^{-1} \|\varphi_\zeta * \psi_r\|_2^2}. \quad (42)$$

• **Variance of $\tilde{U}_{\mathrm{id}}$**

Using the second-order Taylor approximation of $G(\widehat{\lambda})$ around $\lambda_0$ gives

$$\mathrm{Var}[\tilde{U}_{\mathrm{id}}] \leq 4\,\mathrm{Var}[G(\lambda_0)] + 4\,\mathrm{Var}[(\widehat{\lambda} - \lambda_0)\, G'(\lambda_0)] + \frac{1}{2}\,\mathrm{Var}[(\widehat{\lambda} - \lambda_0)^2\, G''(\check{\lambda})]$$

$$\leq Q_0(n^{-2}\zeta^d + n^{-1}\|\varphi_\zeta * \psi_r\|_2^2) + 4\mathbb{E}[(\widehat{\lambda} - \lambda_0)^2 \{G'(\lambda_0)\}^2] + \frac{1}{2}\mathbb{E}[(\widehat{\lambda} - \lambda_0)^4 \{G''(\check{\lambda})\}^2]$$

$$\leq Q_0(n^{-2}\zeta^d + n^{-1}\|\varphi_\zeta * \psi_r\|_2^2) + 4\sqrt{\mathbb{E}[(\widehat{\lambda} - \lambda_0)^4]\,\mathbb{E}[\{G'(\lambda_0)\}^4]} + \frac{1}{2}\sqrt{\mathbb{E}[(\widehat{\lambda} - \lambda_0)^8]\,\mathbb{E}[\{G''(\check{\lambda})\}^4]}$$

$$\leq Q_0(n^{-2}\zeta^d + n^{-1}\|\varphi_\zeta * \psi_r\|_2^2) + Q_1\sqrt{n^{-2}\,\mathbb{E}[\{G'(\lambda_0)\}^4]} + Q_1\sqrt{n^{-4}\,\mathbb{E}[\{G''(\check{\lambda})\}^4]}. \quad (43)$$

Note that the second inequality can be proved following similar lines as in Schrab et al. (2023, Proposition 3), while the last one follows from Lemma 12. Similarly to (40) and (41), we need to control fourth-order moments of some derivatives of $G$. Starting from the term involving $G''$, we can argue similarly to (40) to show $\mathbb{E}[\{G''(\check{\lambda})\}^4] \leq Q_1 n^{-8}\,\mathbb{E}[\{\sum_{k_1 \neq k_2} |k_\zeta(Z_{k_1}, Z_{k_2})|\}^4] \leq Q_1 n^{-4}\,\mathbb{E}[\sum_{k_1 \neq k_2} \sum_{k_3 \neq k_4} k_\zeta^2(Z_{k_1}, Z_{k_2}) k_\zeta^2(Z_{k_3}, Z_{k_4})]$. It is now just a matter of counting what is the contribution of each term in the sum, depending on how many indices are shared. In this regard, observe that we have $\mathcal{O}(n^4)$ terms like $\mathbb{E}[k_\zeta^2(Z_1, Z_2) k_\zeta^2(Z_3, Z_4)] = \mathbb{E}[k_\zeta^2(Z_1, Z_2)]\mathbb{E}[k_\zeta^2(Z_3, Z_4)] \leq M^2 \kappa_2^2(d) \zeta^{2d}$, $\mathcal{O}(n^2)$ terms like $\mathbb{E}[k_\zeta^4(Z_1, Z_2)] \leq M \kappa_4(d) \zeta^{3d}$ and $\mathcal{O}(n^3)$ terms like $\mathbb{E}[k_\zeta^2(Z_1, Z_2) k_\zeta^2(Z_1, Z_3)] = \mathbb{E}[\mathbb{E}[k_\zeta^2(Z_1, Z_2) \mid Z_1]\,\mathbb{E}[k_\zeta^2(Z_1, Z_3) \mid Z_1]] \leq M^2 \kappa_2^2(d) \zeta^{2d}$. Since $\zeta^d/n^2 = n^{-\frac{8s}{4s+d}} \leq 1$

for our specific choice of $\zeta$, this suffices to show that $\sqrt{n^{-4}\,\mathbb{E}[\{G''(\check{\lambda})\}^4]} \leq Q_0 n^{-2}\zeta^d$.

As for the term involving the first derivative of $G$, we have $\mathbb{E}[\{G'(\lambda_0)\}^4] \leq 8\mathbb{E}[\{G'(\lambda_0) - \mathbb{E}G'(\lambda_0)\}^4] + 8\{\mathbb{E}G'(\lambda_0)\}^4 \leq 8\mathbb{E}[\{G'(\lambda_0) - \mathbb{E}G'(\lambda_0)\}^4] + Q_1\|\varphi_\zeta * \psi_r\|_2^4$. Hence, it just remains to bound the first term on the right-hand side. Define $h_{XX}(X_1, X_2) := b_{XX}(X_1, X_2)k_\zeta(X_1, X_2) - \mathbb{E}[b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)]$ and similarly for $h_{YY}$ and $h_{XY}$. We thus have

$$
\mathbb{E}[\{G'(\lambda_0) - \mathbb{E}G'(\lambda_0)\}^4]
$$

$$
= \mathbb{E}\left[\left\{\frac{1}{n(n-1)}\sum_{i_1 \neq i_2} h_{XX}(X_{i_1}, X_{i_2}) + \frac{1}{m(m-1)}\sum_{j_1 \neq j_2} h_{YY}(Y_{j_1}, Y_{j_2}) - \frac{2}{nm}\sum_{i \neq j} h_{XY}(X_i, Y_j)\right\}^4\right]
$$

$$
\leq 32\mathbb{E}\left[\left\{\frac{1}{n(n-1)}\sum_{i_1 \neq i_2} h_{XX}(X_{i_1}, X_{i_2})\right\}^4\right] + 32\mathbb{E}\left[\left\{\frac{1}{m(m-1)}\sum_{j_1 \neq j_2} h_{YY}(Y_{j_1}, Y_{j_2})\right\}^4\right]
$$

$$
+ 32\mathbb{E}\left[\left\{\frac{1}{nm}\sum_{i \neq j} h_{XY}(X_i, Y_j)\right\}^4\right]. \tag{44}
$$

These three terms can be bounded by almost identical arguments so we focus on the first. We expand it as $n^{-4}(n-1)^{-4}\sum_{i_1 \neq i_2}\sum_{i_3 \neq i_4}\sum_{i_5 \neq i_6}\sum_{i_7 \neq i_8}\mathbb{E}[h_{XX}(X_{i_1}, X_{i_2})h_{XX}(X_{i_3}, X_{i_4})h_{XX}(X_{i_5}, X_{i_6})h_{XX}(X_{i_7}, X_{i_8})]$ and use a combinatorial argument to derive an upper bound. In this regard, we have $\mathcal{O}(n^8)$ terms with all distinct indices and $\mathcal{O}(n^7)$ terms with seven distinct indices, but they do not contribute to the sum since their expectations are zero. This is due to the independence among the $X$'s and the fact that $\mathbb{E}[h_{XX}(X_1, X_2)] = 0$. Moreover, we have

$$
\mathbb{E}[h_{XX}(X_{i_1}, X_{i_2})h_{XX}(X_{i_3}, X_{i_4})h_{XX}(X_{i_5}, X_{i_6})h_{XX}(X_{i_7}, X_{i_8})] \leq \mathbb{E}[h_{XX}^4(X_1, X_2)]
$$

$$
= \mathbb{E}[\{b_{XX}(X_1, X_2)k_\zeta(X_1, X_2) - \mathbb{E}[\{b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)]\}^4] \leq 16\mathbb{E}[b_{XX}^4(X_1, X_2)k_\zeta^4(X_1, X_2)]
$$

$$
\leq Q_1\mathbb{E}[k_\zeta^4(X_1, X_2)] \leq Q_1 M\kappa_4(d)\zeta^{3d},
$$

where in first inequality we used the Cauchy-Schwarz inequality, while in the second inequality we used the fact that for $X, X'$ independent and identically distributed we have $\mathbb{E}[(X - \mathbb{E}X)^4] \leq 16\mathbb{E}[X^4]$. Based on this, the contribution of the $\mathcal{O}(n^4)$ terms with four or less distinct indices is bounded above by $n^{-4}\zeta^{3d}$. It remains to bound those terms in which there are five or six different indices. As for the latter case, the only non-zero terms are of the form

$$
\mathbb{E}[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)h_{XX}(X_4, X_5)h_{XX}(X_4, X_6)]
$$

$$
= \mathbb{E}[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)]\,\mathbb{E}[h_{XX}(X_4, X_5)h_{XX}(X_4, X_6)] = \mathbb{E}^2[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)]
$$

$$
= \mathbb{E}^2[\mathbb{E}[h_{XX}(X_1, X_2) \mid X_1]\mathbb{E}[h_{XX}(X_1, X_3) \mid X_1]] \leq \mathbb{E}^2[\mathbb{E}^2[h_{XX}(X_1, X_2) \mid X_1]]
$$

$$
= \mathbb{E}^2[\mathbb{E}^2[\{b_{XX}(X_1, X_2)k_\zeta(X_1, X_2) - \mathbb{E}[b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)]\} \mid X_1]]
$$

$$
\leq 8\mathbb{E}^2[\mathbb{E}^2[b_{XX}(X_1, X_2)k_\zeta(X_1, X_2) \mid X_1]] + 8\mathbb{E}^4[b_{XX}(X_1, X_2)k_\zeta(X_1, X_2)]
$$

$$
\leq Q_1\mathbb{E}^2[\mathbb{E}^2[|k_\zeta(X_1, X_2)| \mid X_1]] + Q_1\mathbb{E}^4[|k_\zeta(X_1, X_2)|] \leq Q_1 M^4\kappa_1^4(d),
$$

and since there are $\mathcal{O}(n^6)$ of these terms, their contribution is of the order $n^{-2}$. Finally, when there are exactly five distinct indices we can have three different typical terms:

1. $\mathbb{E}[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)h_{XX}(X_1, X_4)h_{XX}(X_1, X_5)] = \mathbb{E}[\mathbb{E}^4[h_{XX}(X_1, X_2) \mid X_1]] \leq Q_1 M^4 \kappa_1^4(d);$

2. $\begin{aligned} \mathbb{E}[h_{XX}^2(X_1, X_2)h_{XX}(X_3, X_4)h_{XX}(X_3, X_5)] &= \mathbb{E}[h_{XX}^2(X_1, X_2)]\,\mathbb{E}[h_{XX}(X_3, X_4)h_{XX}(X_3, X_5)] \\ &= \mathbb{E}[h_{XX}^2(X_1, X_2)]\,\mathbb{E}[\mathbb{E}^2[h_{XX}(X_3, X_4) \mid X_3]] \leq Q_1 M^3 \kappa_1^2(d)\kappa_2(d)\zeta^d; \end{aligned}$

3. $\begin{aligned} &\mathbb{E}[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)h_{XX}(X_1, X_4)h_{XX}(X_2, X_5)] \\ &\quad = \mathbb{E}[\mathbb{E}[h_{XX}(X_1, X_2)h_{XX}(X_1, X_3)h_{XX}(X_1, X_4)h_{XX}(X_2, X_5) \mid X_1, X_2]] \\ &\quad \leq \mathbb{E}[|h_{XX}(X_1, X_2)| \cdot \mathbb{E}^2[|h_{XX}(X_1, X_3)| \mid X_1] \cdot \mathbb{E}[|h_{XX}(X_2, X_5)| \mid X_2]] \leq Q_1 M^4 \kappa_1^4(d); \end{aligned}$

There are $\mathcal{O}(n^5)$ of these terms, hence their contribution is of the order $\zeta^d n^{-3}$. Overall, since $\zeta^d/n^2 \leq 1$ for our specific choice of $\zeta$, we thus have that $\sqrt{n^{-2}\mathbb{E}[\{G'(\lambda_0)\}^4]} \leq Q_0\left(n^{-2}\zeta^d + n^{-1}\|\varphi_\zeta * \psi_r\|_2^2\right)$, which further shows that

$$\mathrm{Var}[U_{\mathrm{id}}] \leq Q_0 \left\{ \frac{\zeta^d}{n^2} + \frac{\|\varphi_\zeta * \psi_r\|_2^2}{n} \right\}$$

when combined with (37), (43), and (44). This concludes the analysis for the first two moments of the non-permuted sample.

- **Relating the squared Shifted-MMD to the $L^2$ distance**

Applying the previous argument with (27), (34) and (42) and using the fact that $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$ for $x, y \geq 0$ shows that a uniform control of the Type-I and Type-II errors is possible whenever

$$\mathrm{MMD}_{r,k_\zeta}^2(f, g) \geq Q_0 \left\{ \frac{\zeta^{d/2}}{n} + \frac{\|\varphi_\zeta * \psi_r\|_2}{\sqrt{n}} \right\}.$$

We now conclude the proof by relating the squared Shifted-MMD metric to the $L^2$ distance defined in (13), thus providing an upper bound on $\rho_r^*$. We have

$$\begin{aligned} \mathrm{MMD}_{r,k_\zeta}^2(f, g) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \varphi_\zeta(x-y)\psi_r(x)\psi_r(y)dxdy = \int_{\mathbb{R}^d} \psi_r(x)(\varphi_\zeta * \psi_r)(x)dx \\ &= \langle \psi_r, \varphi_\zeta * \psi_r \rangle_2 = \frac{1}{2}\left(\|\psi_r\|_2^2 + \|\varphi_\zeta * \psi_r\|_2^2 - \|\psi_r - \varphi_\zeta * \psi_r\|_2^2\right), \end{aligned}$$

hence an equivalent sufficient condition to bound the total error is given by

$$\|\psi_r\|_2^2 \geq \|\psi_r - \varphi_\zeta * \psi_r\|_2^2 + Q_0 \frac{\zeta^{d/2}}{n} + Q_0 \frac{\|\varphi_\zeta * \psi_r\|_2}{\sqrt{n}} - \|\varphi_\zeta * \psi_r\|_2^2.$$

This can be further simplified to just $\|\psi_r\|_2^2 \geq \|\psi_r - \varphi_\zeta * \psi_r\|_2^2 + Q_0\,n^{-1}\zeta^{d/2}$ in light of the fact that

$$\sqrt{\frac{Q_0^2\|\varphi_\zeta * \psi_r\|_2^2}{n}} - \|\varphi_\zeta * \psi_r\|_2^2 \leq \frac{Q_0^2}{n} + \|\varphi_\zeta * \psi_r\|_2^2 - \|\varphi_\zeta * \psi_r\|_2^2 \leq \frac{Q_0^2\zeta^{d/2}}{n},$$

as $\zeta \geq 1$ and $\sqrt{xy} \leq x + y$ for $x, y \geq 0$. Furthermore, we can argue exactly as in Schrab et al. (2023, Theorem 6) and show that if $\psi_r \in \mathcal{S}_d^s(L)$ we have $\|\psi_r - \varphi_\zeta * \psi_r\|_2^2 \leq Q_3\|\psi_r\|_2^2 + Q_4\,\zeta^{-2s}$, for some constants

$Q_3 \in (0,1)$ and $Q_4 \equiv Q_4(d,s,L) > 0$. This shows that for $\zeta = n^{\frac{2}{4s+d}}$ there exists a constant $C_r \equiv C_r(c, C, d, \tau, M, s, L, \alpha, \beta)$ such that

$$\rho_r^* \leq C_r \sqrt{\frac{\zeta^{d/2}}{n} + \zeta^{-2s}} = C_r \, n^{-\frac{2s}{4s+d}},$$

and completes the proof. $\qquad\square$

**Lemma 13.** *Fix subsets $\bar{I} \subseteq [n]$ and $\bar{J} \subseteq \{n+1, \ldots, n+m\}$, and modify Step 3 of Algorithm 1 to be: Sample a vector of couples $\tau_t = \{(i_1^t, j_1^t), \ldots, (i_K^t, j_K^t)\}$ such that $(i_1^t, \ldots, i_K^t)$ are sampled uniformly and without replacement from $[n] \setminus \bar{I}$, and $(j_1^t, \ldots, j_K^t)$ are sampled uniformly and without replacement from $\{n+1, \ldots, n+m\} \setminus \bar{J}$. Keep all the other steps the same. Then this algorithm still has (4) as its stationary distribution.*

*Proof.* The key observation to make here is that all the steps that led to (18) in the proof of Proposition 2 remain valid in this other setting. The only difference lies in the fact that the Markov chain associated with this new algorithm is not irreducible, and hence it will not converge to (4) if we let it run long enough. Nonetheless, (4) is still a stationary distribution, and we will now show this by proving a detailed balance condition. Let $K = \min\{n - \#\bar{I}, m - \#\bar{J}\}$ and let $\tilde{\mathcal{P}}$ be the set of all $K$ couples of the form $\{(i_1, j_1), \ldots, (i_K, j_K)\}$ such that $(i_1, \ldots, i_K)$ contains distinct elements from $[n] \setminus \bar{I}$, and $(j_1, \ldots, j_K)$ contains distinct elements from $\{n+1, \ldots, n+m\} \setminus \bar{J}$. For all $t \in \mathbb{N}_+$ and all permutations $p, p'$, we have

$$\mathbb{P}\{P_t = p' \mid P_{t-1} = p\} = \frac{1}{|\tilde{\mathcal{P}}|} \sum_{\tau \in \tilde{\mathcal{P}}} \mathbb{P}\{P_t = p' \mid P_{t-1} = p, \tau_t = \tau\},$$

since at each time $t$ this new algorithm draws $\tau_t \in \tilde{\mathcal{P}}$ uniformly at random. Next, given $\tau_t = \tau$ and $P_{t-1} = p$, it must be the case that $P_t$ satisfies $P_t \sim_\tau p$, since this new algorithm still uses Steps 4-5 of Algorithm 1. Arguing as in (17) gives

$$\frac{\mathbb{P}\{P_t = p' \mid P_{t-1} = p, \tau_t = \tau\}}{\mathbb{P}\{P_t = p'' \mid P_{t-1} = p, \tau_t = \tau\}} = \frac{\mathbb{P}\{P = p'\}}{\mathbb{P}\{P = p''\}},$$

which implies that

$$\mathbb{P}\{P_t = p' \mid P_{t-1} = p\} = \frac{1}{|\tilde{\mathcal{P}}|} \sum_{\tau \in \tilde{\mathcal{P}}} \frac{\mathbb{1}\{p' \sim_\tau p\} \cdot \mathbb{P}\{P = p'\}}{\sum_{p''} \mathbb{1}\{p'' \sim_\tau p\} \cdot \mathbb{P}\{P = p''\}}.$$

This concludes the proof by analogous calculations to those in (18). $\qquad\square$

*Proof of Theorem 10.* For simplicity, we assume $n = m$ throughout the proof. The more general case $m \leq n \leq \tau m$ corresponds to a simpler problem, as it involves a larger sample size; thus, the lower bound derived here remains valid in that setting. For varying densities $p, q$ on $\mathbb{R}^d$ define the set

$$\tilde{\mathcal{S}}_\theta^r(\rho) := \mathcal{S}_\theta^r(\rho) \cap \left\{ (f \equiv f_p, g \equiv g_q) : f_p = \gamma_p \frac{p}{r}\left(1 + r\frac{\gamma_q}{\gamma_p}\right), g_q = \gamma_q q\left(1 + r\frac{\gamma_q}{\gamma_p}\right) \right\} \subseteq \mathcal{S}_\theta^r(\rho), \qquad (45)$$

where $\gamma_p = \frac{\sqrt{B}}{A\sqrt{B}+\sqrt{A}}$ and $\gamma_q = \frac{\sqrt{A}}{A\sqrt{B}+\sqrt{A}}$, with $A = \int_{\mathbb{R}^d} p(x)/r(x)dx$ and $B = \int_{\mathbb{R}^d} q(x)r(x)dx$. One can easily check that $\int_{\mathbb{R}^d} f_p(x)dx = \int_{\mathbb{R}^d} g_q(x)dx = 1$ for this specific choice of $\gamma_p$ and $\gamma_q$. As a result, for any

test $\varphi \in \Psi(\alpha)$, and for all prior distributions $\mu_0, \mu_1$ supported on $H_0$ and $\tilde{\mathcal{S}}_\theta^r(\rho)$, respectively, Equation (45) shows that we can bound the total error probability as

$$
\begin{aligned}
\alpha + \sup_{(f,g)\in\mathcal{S}_\theta^r(\rho)} \mathbb{E}_P(1-\varphi) &\geq \sup_{g \propto rf} \mathbb{E}_P \varphi + \sup_{(f,g)\in\mathcal{S}_\theta^r(\rho)} \mathbb{E}_P(1-\varphi) \\
&\geq \sup_{g \propto rf} \mathbb{E}_P \varphi + \sup_{(f,g)\in\tilde{\mathcal{S}}_\theta^r(\rho)} \mathbb{E}_P(1-\varphi) \geq \mathbb{E}_{\mu_0}\{\mathbb{E}_P \varphi\} + \mathbb{E}_{\mu_1}\{\mathbb{E}_P(1-\varphi)\} \geq 1 - \mathrm{TV}(\mathbb{E}_{\mu_0}P, \mathbb{E}_{\mu_1}P),
\end{aligned}
\tag{46}
$$

where we recall that $P = P_f^{\otimes n} \otimes P_g^{\otimes n}$. This demonstrates that controlling the total variation distance above is sufficient to obtain a lower bound on $\rho_r^*$. We proceed to do so for specific choices of $\mu_0$ and $\mu_1$, using a classical perturbation-based method originating from Ingster (1987) and recently employed in two-sample and independence testing problems in Albert et al. (2022) and Li and Yuan (2024). Start by considering $q_0(x) = q_1(x) = p_0(x) = \mathbb{1}\{x \in [0,1]^d\}$ and define $A_i = \int_{\mathbb{R}^d} p_i(x)/r(x)dx$ and $B_i = \int_{\mathbb{R}^d} q_i(x)r(x)dx$ for $i \in \{0,1\}$. We can assume without loss of generality that $\int_{[0,1]^d} r(x)dx = 1$ so that $B_0 = B_1 = 1$; otherwise, since the problem is scale invariant, we might replace $r$ with $r/\int_{[0,1]^d} r(x)dx$ without affecting the minimax separation. Finally, define $p_1$ to be a perturbation of $p_0$ of the following form. Let $\tilde{M}$ be as in (53), fix

$$
B_n^{1/d} = \left\lfloor \left( \frac{\sqrt{c}(2\pi)^{d/2}L}{2\sqrt{2(1+c)}\,\tilde{M}\rho} \right)^{1/s} \right\rfloor \qquad \text{and} \qquad \delta_n = \sqrt{\frac{8(1+c)}{c}} \rho\, B_n^{-1/2},
$$

and consider $\{\phi_1, \dots, \phi_{B_n}\}$ as in (52) in the proof of Lemma 14, i.e. an orthonormal set of functions in $L^2(\mathbb{R}^d)$ whose supports are disjoint and contained in $[0,1]^d$, and satisfy

$$
\int_{\mathbb{R}^d} \phi_k(x)dx = \int_{\mathbb{R}^d} \frac{\phi_k(x)}{r(x)}dx = 0 \quad \text{for all } k \in [B_n].
\tag{47}
$$

It is convenient to recall that the $\phi_k$'s are of the form $\phi_k(x) = \frac{B_n^{1/2}}{\|\phi_{0,k}\|_2} \phi_{0,k}\big(B_n^{1/d}\{x - x_k^0\}\big)$, where $x_k^0$ is the lower-left corner of their support, and the $\phi_{0,k}$'s satisfy $\max_{k\in[B_n]} \left\{ \frac{\|\phi_{0,k}\|_{\mathcal{S}_d^s}}{\|\phi_{0,k}\|_2} \vee \frac{\|\phi_{0,k}\|_\infty}{\|\phi_{0,k}\|_2} \right\} \leq \tilde{M}$, with $\|\cdot\|_{\mathcal{S}_d^s}$ defined in (49). Based on this, we define

$$
p_1(x) \equiv p_{1,a}(x) = p_0(x) + \delta_n \sum_{k=1}^{B_n} a_k \, \phi_k(x),
\tag{48}
$$

where $a = (a_1, \dots, a_{B_n})$ is a collection of i.i.d. Rademacher random variables, meaning that $\mathbb{P}\{a_k = 1\} = \mathbb{P}\{a_k = -1\} = 1/2$ for all $k \in [B_n]$.

With these definitions in mind, let $f_i := f_{p_i}$ and $g_i := g_{q_i}$ for $i \in \{0,1\}$, with $f_p, g_p$ as in (45). Now, it is clear that $(f_0, g_0)$ satisfy the null, thus we just need to check that $(f_1, g_1) \in \tilde{\mathcal{S}}_\theta^r(\rho)$ for all $a \in \{\pm 1\}^{B_n}$, which is required to ensure that the distribution $\mu_1$ that assigns equal probability to each of them is indeed supported on $\tilde{\mathcal{S}}_\theta^r(\rho)$. We may assume that $\|f_1\|_\infty \vee \|g_1\|_\infty \leq M$; otherwise, it suffices to choose $p_0(x) = \mathbb{1}\{x \in [0,u]^d\}$ for $u \geq 1$ sufficiently large and construct a perturbed version of it as in (48). This is clearly sufficient for $g_1$, since the prefactor of $q$ in the definition of $g_q$ in (45) depends on $r$, which is uniformly bounded. A similar

argument applies to $f_1$, taking into account that the magnitudes of the bumps are bounded as

$$\|\delta_n \sum_{k=1}^{B_n} a_k \phi_k\|_\infty \leq \delta_n \max_{k=1}^{B_n} \|\phi_k\|_\infty \leq \delta_n B_n^{1/2} \max_{k=1}^{B_n} \frac{\|\phi_{0,k}\|_\infty}{\|\phi_{0,k}\|_2} \leq \tilde{M} \delta_n B_n^{1/2} \lesssim \rho \lesssim 1.$$

As for the conditions involving $\psi_r$, observe that

$$\psi_r = \frac{\lambda_0 r f_1 - g_1}{1 + \lambda_0 r} = \frac{1 + r \gamma_{q_1}/\gamma_{p_1}}{1 + \lambda_0 r}(\lambda_0 \gamma_{p_1} p_1 - \gamma_{q_1} q_1) = \gamma_{q_1}(p_1 - q_1),$$

since $\lambda_0 = \gamma_{q_1}/\gamma_{p_1}$ satisfies $\int_{\mathbb{R}^d} \psi_r(x)dx = 0$. Hence we need to verify that $\gamma_{q_1}\|p_1-q_1\|_2 > \rho$ and $\gamma_{q_1}(p_1-q_1) \in \mathcal{S}_d^s(L)$. Start by noticing that for all $a \in \{\pm 1\}^{B_n}$ we have

$$1 \geq \gamma_{q_1}^2 = (\sqrt{A_1} + 1)^{-2} \geq (2A_1 + 2)^{-1} \geq \frac{c}{2(1+c)},$$

as

$$A_1 = \int_{\mathbb{R}^d} \frac{p_1(x)}{r(x)}dx = \int_{\mathbb{R}^d} \frac{p_0(x) + \delta_n \sum_{k=1}^{B_n} a_k \phi_k(x)}{r(x)}dx \stackrel{(47)}{=} \int_{\mathbb{R}^d} \frac{p_0(x)}{r(x)}dx \leq \frac{1}{c}.$$

Then, as for the condition involving the $L^2$-norm of $\psi_r$, we have

$$\gamma_{q_1}^2\|p_1 - q_1\|_2^2 = \gamma_{q_1}^2\left\|\delta_n \sum_{k=1}^{B_n} a_k \phi_k\right\|_2^2 = \gamma_{q_1}^2 \delta_n^2\left\|\sum_{k=1}^{B_n} a_k \phi_k\right\|_2^2 = \gamma_{q_1}^2 \delta_n^2 \int_{\mathbb{R}^d}\left(\sum_{k=1}^{B_n} a_k \phi_k(x)\right)^2 dx$$

$$= \gamma_{q_1}^2 \delta_n^2 \sum_{k=1}^{B_n} a_k^2 \int_{\mathbb{R}^d} \phi_k^2(x)dx = \gamma_{q_1}^2 \delta_n^2 B_n \geq \frac{c}{2(1+c)}\delta_n^2 B_n = 4\rho^2 > \rho^2$$

for our particular choice of $\delta_n$ and $B_n$. As for the smoothness condition, define for notational convenience the norm

$$\|p\|_{\mathcal{S}_d^s}^2 := \int_{\mathbb{R}^d} \|\xi\|_2^{2s}|\widehat{p}(\xi)|^2 \, d\xi, \tag{49}$$

so that $\mathcal{S}_d^s(L) = \{p \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d) : \|p\|_{\mathcal{S}_d^s}^2 \leq (2\pi)^d L^2\}$. Furthermore, since the iterated Laplacian $(-\Delta)^s$ of order $s \in \mathbb{N}_+$ is the Fourier multiplier with symbol $\|\xi\|_2^{2s}$, Plancherel's theorem gives

$$\int_{\mathbb{R}^d} \|\xi\|_2^{2s} \widehat{\phi_1}(\xi) \overline{\widehat{\phi_2}(\xi)} \, d\xi = \int_{\mathbb{R}^d} (-\Delta)^s \phi_1(x) \overline{\phi_2(x)} \, dx = 0,$$

since $(-\Delta)^s \phi_1$ is a combination of derivatives of $\phi_1$ of order $2s$, and $\phi_1, \phi_2$ have disjoint supports. This implies that

$$\gamma_{q_1}^2\|p_1 - q_1\|_{\mathcal{S}_d^s}^2 \leq \|p_1 - q_1\|_{\mathcal{S}_d^s}^2 = \left\|\delta_n \sum_{k=1}^{B_n} a_k \phi_k\right\|_{\mathcal{S}_d^s}^2 = \delta_n^2\left\|\sum_{k=1}^{B_n} a_k \phi_k\right\|_{\mathcal{S}_d^s}^2 = \delta_n^2 \int_{\mathbb{R}^d} \|\xi\|_2^{2s}\left|\sum_{k=1}^{B_n} a_k \widehat{\phi_k}(\xi)\right|^2 d\xi$$

$$
\leq \delta_n^2 B_n \max_{k=1}^{B_n} \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left| \widehat{\phi_k}(\xi) \right|^2 \, \mathrm{d}\xi = \delta_n^2 B_n \max_{k=1}^{B_n} \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left| \frac{B_n^{1/2}}{\|\phi_{0,k}\|_2} \widehat{\phi_{0,k}(B_n^{1/d}\{\cdot - x_k^0\})}(\xi) \right|^2 \, \mathrm{d}\xi
$$

$$
= \delta_n^2 B_n \max_{k=1}^{B_n} \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left| \frac{B_n^{-1/2}}{\|\phi_{0,k}\|_2} e^{-i\langle \xi, x_k^0 \rangle} \widehat{\phi_{0,k}}(\xi/B_n^{1/d}) \right|^2 \, \mathrm{d}\xi = \max_{k=1}^{B_n} \frac{\delta_n^2}{\|\phi_{0,k}\|_2^2} \int_{\mathbb{R}^d} \|\xi\|_2^{2s} \left| \widehat{\phi_{0,k}}(\xi/B_n^{1/d}) \right|^2 \, \mathrm{d}\xi
$$

$$
= \max_{k=1}^{B_n} \frac{\delta_n^2}{\|\phi_{0,k}\|_2^2} \int_{\mathbb{R}^d} \|B_n^{1/d}\xi\|_2^{2s} \left| \widehat{\phi_{0,k}}(\xi) \right|^2 B_n \, \mathrm{d}\xi = \delta_n^2 B_n^{\frac{2s+d}{d}} \max_{k=1}^{B_n} \frac{\|\phi_{0,k}\|_{\dot{\mathcal{S}}_d^s}^2}{\|\phi_{0,k}\|_2^2} \leq \tilde{M}^2 \delta_n^2 B_n^{\frac{2s+d}{d}} \leq (2\pi)^d L^2
$$

for our particular choice of $\delta_n$ and $B_n$.

It remains to control the total variation in (46) for this specific choice of $\mu_1, \delta_n$ and $B_n$, and assess for which values of $\rho$ we can bound it above by $1 - \alpha - \beta$. Writing $f_1 \equiv f_{1,a}$ and $g_1 \equiv g_{1,a}$ to highlight their dependence on $a \in \{\pm 1\}^d$ through $p_1$, and using $\chi^2$ for the chi-square divergence, we look at

$$
4\,\mathrm{TV}^2(P_{f_0}^{\otimes n} \otimes P_{g_0}^{\otimes n}, \mathbb{E}_a\{P_{f_1}^{\otimes n} \otimes P_{g_1}^{\otimes n}\}) = 4\,\mathrm{TV}^2 \left( P_{f_0}^{\otimes n} \otimes P_{g_0}^{\otimes n}, 2^{-B_n} \sum_{a \in \{\pm 1\}^d} P_{f_{1,a}}^{\otimes n} \otimes P_{g_{1,a}}^{\otimes n} \right)
$$

$$
\leq \chi^2 \left( P_{f_0}^{\otimes n} \otimes P_{g_0}^{\otimes n}, 2^{-B_n} \sum_{a \in \{\pm 1\}^d} P_{f_{1,a}}^{\otimes n} \otimes P_{g_{1,a}}^{\otimes n} \right)
$$

$$
= 2^{-2B_n} \sum_{a_1, a_2 \in \{\pm 1\}^d} \int_{\mathbb{R}^{2nd}} \left( \prod_{i=1}^n \frac{f_{1,a_1}(x_i)}{f_0(x_i)} \prod_{j=1}^n \frac{g_{1,a_1}(y_j)}{g_0(y_j)} \right) \left( \prod_{i=1}^n \frac{f_{1,a_2}(x_i)}{f_0(x_i)} \prod_{j=1}^n \frac{g_{1,a_2}(y_j)}{g_0(y_j)} \right) dP_0
$$

$$
= 2^{-2B_n} \sum_{a_1, a_2 \in \{\pm 1\}^d} \left( \prod_{i=1}^n \int_{\mathbb{R}^d} \frac{f_{1,a_1}(x_i)}{f_0(x_i)} \frac{f_{1,a_2}(x_i)}{f_0(x_i)} f_0(x_i) dx_i \right) \left( \prod_{j=1}^n \int_{\mathbb{R}^d} \frac{g_{1,a_1}(y_j)}{g_0(y_j)} \frac{g_{1,a_2}(y_j)}{g_0(y_j)} g_0(y_j) dy_j \right)
$$

$$
= 2^{-2B_n} \sum_{a_1, a_2 \in \{\pm 1\}^d} \left( \int_{\mathbb{R}^d} \frac{f_{1,a_1}(x_1)}{f_0(x_1)} \frac{f_{1,a_2}(x_1)}{f_0(x_1)} f_0(x_1) dx_1 \right)^n \left( \int_{\mathbb{R}^d} \frac{g_{1,a_1}(y_1)}{g_0(y_1)} \frac{g_{1,a_2}(y_1)}{g_0(y_1)} g_0(y_1) dy_1 \right)^n \quad (50)
$$

where we set $dP_0 = \prod_{i=1}^n f_0(x_i) \prod_{j=1}^n g_0(y_j) dx dy$. We now focus on controlling the integrals in the last display. As for the latter, Equation (47) implies that $A_0 = A_1$, hence $\gamma_{p_1} = \gamma_{p_0}$ and $\gamma_{q_1} = \gamma_{q_0}$. This gives $\frac{g_{1,a}(y_1)}{g_0(y_1)} = \mathbb{1}\{x \in [0,1]^d\}$ for all $a \in \{\pm 1\}^d$, and further shows that the second integral in (50) is equal to one. As for the other one, similar calculations show that

$$
\int_{\mathbb{R}^d} \frac{f_{1,a_1}(x)}{f_0(x)} \frac{f_{1,a_2}(x)}{f_0(x)} f_0(x) dx = \int_{\mathbb{R}^d} \left( 1 + \delta_n \sum_{k=1}^{B_n} a_{1,k} \frac{\phi_k(x)}{p_0(x)} \right) \left( 1 + \delta_n \sum_{k=1}^{B_n} a_{2,k} \frac{\phi_k(x)}{p_0(x)} \right) f_0(x) dx
$$

$$
= 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx + \delta_n \sum_{k=1}^{B_n} (a_{1,k} + a_{2,k}) \int_{\mathbb{R}^d} \frac{\phi_k(x)}{p_0(x)} f_0(x) dx
$$

$$
= 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx + \delta_n \sum_{k=1}^{B_n} (a_{1,k} + a_{2,k}) \int_{\mathbb{R}^d} \frac{\phi_k(x)}{p_0(x)} \frac{p_0(x)}{r(x)} (\gamma_{p_0} + r\,\gamma_{q_0}) dx
$$

$$
= 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx + \delta_n \sum_{k=1}^{B_n} (a_{1,k} + a_{2,k}) \left\{ \gamma_{p_0} \int_{\mathbb{R}^d} \frac{\phi_k(x)}{r(x)} dx + \gamma_{q_0} \int_{\mathbb{R}^d} \phi_k(x) dx \right\}
$$

$$= 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \tag{51}$$

where the second equality follows from the fact that $\phi_{k_1}$ and $\phi_{k_2}$ have disjoint support when $k_1 \neq k_2$, and the final equality uses condition (47). Combining (50) with (51) then shows

$$4 \operatorname{TV}^2(P_{f_0}^{\otimes n} \otimes P_{g_0}^{\otimes n}, \mathbb{E}_a\{P_{f_1}^{\otimes n} \otimes P_{g_1}^{\otimes n}\}) \le 2^{-2B_n} \sum_{a_1, a_2 \in \{\pm 1\}^d} \left( 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right)^n$$

$$= \mathbb{E}_{a_1, a_2} \left[ \left( 1 + \delta_n^2 \sum_{k=1}^{B_n} a_{1,k} a_{2,k} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right)^n \right] = \mathbb{E}_a \left[ \left( 1 + \delta_n^2 \sum_{k=1}^{B_n} a_k \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right)^n \right]$$

$$\le \mathbb{E}_a \left[ \exp \left\{ n \delta_n^2 \sum_{k=1}^{B_n} a_k \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right\} \right] = \prod_{k=1}^{B_n} \cosh \left( n \delta_n^2 \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right)$$

$$\le \exp \left\{ \frac{1}{2} B_n n^2 \delta_n^4 \left( \overset{B_n}{\underset{k=1}{\max}} \int_{\mathbb{R}^d} \frac{\phi_k^2(x)}{p_0^2(x)} f_0(x) dx \right)^2 \right\} = \exp \left\{ \frac{1}{2} B_n n^2 \delta_n^4 \left( \overset{B_n}{\underset{k=1}{\max}} \int_{\mathbb{R}^d} \frac{\gamma_{p_0} + r(x)\gamma_{q_0}}{r(x)} \phi_k^2(x) dx \right)^2 \right\}$$

$$\le \exp \left\{ \frac{C + c^2}{c^2} B_n n^2 \delta_n^4 \right\},$$

where in the last step we used the fact that $c \le r(\cdot) \le C$ together with $\gamma_{q_0} = (\sqrt{A_0} + 1)^{-1} \le 1$ and $\gamma_{p_0} = A_0^{-1/2} \gamma_{q_0} \le A_0^{-1/2} = \{\int_{\mathbb{R}^d} p_0(x)/r(x)\}^{-1/2} \le \sqrt{C}$. Now, being $B_n n^2 \delta_n^4$ of the order $n^2 \rho^{\frac{4s+d}{s}}$, the previous shows that there exists a constant $c_r = c_r(c, C, \theta, \alpha, \beta)$ for which the previous display is upper bounded by $1 - \alpha - \beta$ whenever $\rho \le c_r n^{-2s/(4s+d)}$. This concludes the proof. $\qquad \square$

**Lemma 14.** *Let $r : \mathbb{R}^d \to \mathbb{R}_+$ be such that $0 < c \le r(x) \le C$ for all $x \in \mathbb{R}^d$. Fix an integer $B_n \ge 1$ such that $B_n^{1/d} \in \mathbb{N}_+$ and write $b_n := B_n^{-1/d}$. Partition $[0,1]^d$ into the $B_n$ disjoint cubes*

$$Q_k = \prod_{j=1}^d [i_j b_n, (i_j + 1) b_n), \qquad \text{with } k = (i_1, \ldots, i_d) \in \{0, \ldots, b_n^{-1} - 1\}^d,$$

*and denote the lower–left corner of $Q_k$ by $x_k^0$. There exist functions $\{\phi_{0,1}, \ldots, \phi_{0,B_n}\}$ supported on $[0,1]^d$ satisfying the following conditions:*

(i) *For all $k \in [B_n]$*

$$\phi_k(x) = \frac{B_n^{1/2}}{\|\phi_{0,k}\|_2} \phi_{0,k}(B_n^{1/d}\{x - x_k^0\})$$

*is $C^\infty(\mathbb{R}^d)$ and satisfies*

$$\int_{\mathbb{R}^d} \phi_k(x) \, dx = \int_{\mathbb{R}^d} \frac{\phi_k(x)}{r(x)} \, dx = 0, \qquad \|\phi_k\|_2 = 1, \qquad \operatorname{supp} \phi_k \subseteq Q_k;$$

(ii) *There exist a constant $\tilde{M} > 0$ such that*

$$\overset{B_n}{\underset{k=1}{\max}} \left\{ \frac{\|\phi_{0,k}\|_{\mathcal{S}_d^s}}{\|\phi_{0,k}\|_2} \vee \frac{\|\phi_{0,k}\|_\infty}{\|\phi_{0,k}\|_2} \right\} \le \tilde{M}.$$

*Proof.* We will start by proving part $(i)$. Consider the standard $C^\infty(\mathbb{R})$ bump

$$\eta(t) := \begin{cases} \exp\left(-\frac{1}{1-t^2}\right), & |t| < 1, \\ 0, & |t| \geq 1, \end{cases}$$

and define $f_+(y_1) = \eta(6y_1 - 1)$, $f_-(y_1) = \eta(6y_1 - 3)$ and $f_0(y_1) = \eta(6y_1 - 5)$. It is immediate to see that $f_+$, $f_-$ and $f_0$ are supported on $[0, \frac{1}{3}]$, $[\frac{1}{3}, \frac{2}{3}]$ and $[\frac{2}{3}, 1]$, respectively. Based on this, define three bumps in $\mathbb{R}^d$ by

$$\phi_+(y) = f_+(y_1) \prod_{l=2}^{d} \eta(y_l), \quad \phi_-(y) = f_-(y_1) \prod_{l=2}^{d} \eta(y_l), \quad \phi_0(y) = f_0(y_1) \prod_{l=2}^{d} \eta(y_l)$$

for all $y = (y_1, \ldots, y_d) \in \mathbb{R}^d$, and observe that all three functions are in $C^\infty(\mathbb{R}^d)$ and supported in $[0, \frac{1}{3}] \times [0, 1]^{d-1}$, $[\frac{1}{3}, \frac{2}{3}] \times [0, 1]^{d-1}$ and $[\frac{2}{3}, 1] \times [0, 1]^{d-1}$, respectively. Fix a cube $Q_k$ and write points in it as $x = x_k^0 + b_n y$ with $y \in [0, 1]^d$. Because $r^{-1}$ is bounded on $Q_k$, each of the numbers

$$a_k = \int_{\mathbb{R}^d} \frac{\phi_+(y)}{r(x_k^0 + b_n y)} \, dy, \; b_k = \int_{\mathbb{R}^d} \frac{\phi_-(y)}{r(x_k^0 + b_n y)} \, dy, \; c_k = \int_{\mathbb{R}^d} \frac{\phi_0(y)}{r(x_k^0 + b_n y)} \, dy$$

is finite and strictly positive. If $a_k \neq b_k$, set

$$u_k = \frac{b_k - c_k}{b_k - a_k}, v_k = \frac{a_k - c_k}{b_k - a_k}, w_k = 1 \qquad \text{and} \qquad \phi_{0,k}(y) = u_k \, \phi_+(y) - v_k \, \phi_-(y) - w_k \phi_0(y).$$

If instead $a_k = b_k$ but $a_k \neq c_k$, we switch the roles of $\phi_-$ and $\phi_0$ and apply the same formula. Finally, if $a_k = b_k = c_k$, we simply set $u_k = w_k = 1$ and $v_k = 0$, so thaty $\phi_{0,k}(y) = \phi_+(y) - \phi_0(y)$. A straightforward calculation shows that $\phi_{0,k}$ has zero average both with respect to Lebesgue measure and with respect to the weight $r^{-1}$ evaluated at $x_k^0 + b_n y$:

$$\int_{\mathbb{R}^d} \phi_{0,k}(y) \, dy = \left( \int_{\mathbb{R}^d} \phi_+(y) \, dy \right) (u_k - v_k - w_k) = 0, \qquad \int_{\mathbb{R}^d} \frac{\phi_{0,k}(y)}{r(x_k^0 + b_n y)} \, dy = u_k a_k - v_k b_k - w_k c_k = 0.$$

We now verify that the function

$$\phi_k(x) := \frac{B_n^{1/2}}{\|\phi_{0,k}\|_2} \phi_{0,k}\big(B_n^{1/d}\{x - x_k^0\}\big) \tag{52}$$

satisfies the desired properties. Because $dx = b_n^d dy = B_n^{-1} dy$, the two zero–average identities above translate to the $x$-scale, yielding

$$\int_{\mathbb{R}^d} \phi_k(x) \, dx = \int_{\mathbb{R}^d} \frac{\phi_k(x)}{r(x)} \, dx = 0.$$

Moreover $\|\phi_k\|_2^2 = B_n \|\phi_{0,k}\|_2^{-2} B_n^{-1} \|\phi_{0,k}\|_2^2 = 1$. Finally, the support of $\phi_k$ is contained in $Q_k$, and different cubes do not intersect, hence for $k_1 \neq k_2$ we also have $\langle \phi_{k_1}, \phi_{k_2} \rangle_2 = 0$. This completes the proof of the of first part of the statement.

As for part $(ii)$, set $J_\ell := \|\phi_+\|_\ell = \|\phi_-\|_\ell = \|\phi_0\|_\ell$ for $\ell \in \{\infty, 2, \mathcal{S}_d^s\}$, and observe that these constants

depend only on the shape of the function $\eta$. The disjoint structure of the supports of $\phi_+, \phi_-, \phi_0$ gives

$$
\begin{cases}
\|\phi_{0,k}\|_\infty = J_\infty \max\{|u_k|, |v_k|, |w_k|\} \\
\|\phi_{0,k}\|_2 = J_2 \sqrt{u_k^2 + v_k^2 + w_k^2} \\
\|\phi_{0,k}\|_{\mathcal{S}_d^s} = J_{\mathcal{S}_d^s} \sqrt{u_k^2 + v_k^2 + w_k^2}
\end{cases}
$$

for all $k \in [B_k]$, and implies that

$$
\frac{\|\phi_{0,k}\|_{\mathcal{S}_d^s}}{\|\phi_{0,k}\|_2} = \frac{J_{\mathcal{S}_d^s}}{J_2} \qquad \text{and} \qquad \frac{\|\phi_{0,k}\|_\infty}{\|\phi_{0,k}\|_2} = \frac{J_\infty \max\{|u_k|, |v_k|, |w_k|\}}{J_2 \sqrt{u_k^2 + v_k^2 + w_k^2}} \leq \frac{J_\infty(|u_k| + |v_k| + |w_k|)}{J_2 \sqrt{u_k^2 + v_k^2 + w_k^2}} \leq \frac{\sqrt{3} J_\infty}{J_2}.
$$

Note the denominator is always well-defined for our choices of $(u_k, v_k, w_k)$ as $u_k^2 + v_k^2 + w_k^2 \geq 1$. This shows that the claim in part $(ii)$ holds with

$$
\tilde{M} := \frac{J_{\mathcal{S}_d^s}}{J_2} \vee \frac{\sqrt{3} J_\infty}{J_2} \tag{53}
$$

and concludes the proof of the lemma. $\qquad\square$

## A.3   Proofs for Section 4

*Proof of Proposition 11.* The reason we cannot directly apply Theorem 1 or Proposition 3 is that, although $H_0^{\mathcal{R}}$ holds for a certain $r_\star$, the DRPT procedure now generates $Z^{(1)}, \ldots, Z^{(H)}$ using an approximation $\widehat{r}$ to $r_\star$. To address this mismatch, let $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_n)$, $\tilde{Y} = (\tilde{Y}_1, \ldots, \tilde{Y}_m)$ be such that $\tilde{X} \perp\!\!\!\perp \tilde{Y}$, $\tilde{X}_i \overset{\text{i.i.d.}}{\sim} f$ and $\tilde{Y}_i \overset{\text{i.i.d.}}{\sim} \bar{r} f$. Define $\tilde{Z} = (\tilde{X}, \tilde{Y})$ and let $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}$ be draws of the DRPT based on $\widehat{r}$ when we sample from the values of $\tilde{Z}$ instead of $Z$. That is, for every $h \in [H]$ independently we have

$$
\tilde{Z}^{(h)} = \tilde{Z}_{(\tilde{P}^{(h)})} \quad \text{where} \quad \mathbb{P}\left\{ \tilde{P}^{(h)} = p \mid \tilde{Z}_{()} \right\} \propto \prod_{i \in \{n+1, \ldots, n+m\}} \widehat{r}(\tilde{Z}_{(p(i))}),
$$

where $\tilde{Z}_{()}$ and $\tilde{Z}_{(p)}$ are defined analogously to $Z_{()}$ and $Z_{(p)}$. Next, by comparing to the DRPT sampling mechanism (4), we observe that the $\tilde{Z}^{(h)}$'s, conditional on $\tilde{Z}$, are generated with the same mechanism as the $Z^{(h)}$'s conditional on $Z$. That is, for every $z \in \mathcal{X}^n \times \mathcal{Y}^m$ we have

$$
\left( (\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}) \mid \tilde{Z} = z \right) \overset{d}{=} \left( (Z^{(1)}, \ldots, Z^{(H)}) \mid Z = z \right).
$$

We can verify this also for the exchangeable sampler (Algorithm 2) with a generic parameter $S \geq 1$. We can now use the fact that, if $(V \mid U = u) \overset{d}{=} (V' \mid U' = u)$ for all $u$, then $\text{TV}((U, V), (U', V')) = \text{TV}(U, U')$. It follows that

$$
\text{TV}\left( (\tilde{Z}, \tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}), (Z, Z^{(1)}, \ldots, Z^{(H)}) \right) = \text{TV}(\tilde{Z}, Z) = \text{TV}\left( (\tilde{X}, \tilde{Y}), (X, Y) \right)
$$

$$
= \text{TV}(\tilde{Y}, Y) = \text{TV}\left( \{\bar{r} \cdot f\}^{\otimes m}, \{\bar{r}_\star \cdot f\}^{\otimes m} \right). \tag{54}
$$

We are now in the position to conclude the proof. If we define

$$A_\alpha := \left\{ (z, z^{(1)}, \ldots, z^{(H)}) : \frac{1 + \sum_{h=1}^H \mathbb{1}\{T(z^{(h)}) \geq T(z)\}}{1 + H} \leq \alpha \right\},$$

we can bound the Type-I error as

$$
\begin{aligned}
\mathbb{P}\{p \leq \alpha\} &= \mathbb{P}\{(Z, Z^{(1)}, \ldots, Z^{(H)}) \in A_\alpha\} \\
&\leq \mathbb{P}\{(\tilde{Z}, \tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}) \in A_\alpha\} + \mathrm{TV}\left( (\tilde{Z}, \tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}), (Z, Z^{(1)}, \ldots, Z^{(H)}) \right) \\
&= \mathbb{P}\{(\tilde{Z}, \tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}) \in A_\alpha\} + \mathrm{TV}\left( \{\bar{r} \cdot f\}^{\otimes m}, \{\bar{r}_\star \cdot f\}^{\otimes m} \right) \\
&\leq \alpha + \mathrm{TV}\left( \{\bar{r} \cdot f\}^{\otimes m}, \{\bar{r}_\star \cdot f\}^{\otimes m} \right),
\end{aligned}
$$

where the first inequality follows from the definition of Total Variation distance, the second equality from (54), and the final inequality from the exchangeability of $(\tilde{Z}, \tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)})$. This arises from the fact that the $\tilde{Y}_i$'s are i.i.d. from a distribution proportional to $\hat{r} f$, and the DRPT copies $\tilde{Z}^{(1)}, \ldots, \tilde{Z}^{(H)}$ are generated using the same approximation $\hat{r}$. This argument holds whether the permutations are sampled i.i.d. according to (4) or generated via Algorithm 2, and thus concludes the proof. □

# Appendix B  Analysis of the discrete DRPT

In this section we provide some further insights for the discrete DRPT presented in Section 2.1. For the case $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ (i.e. $J = 1$), we recall that the testing problem (1) is equivalent to

$$H_0 : \frac{g_1}{g_0} = \frac{r_1 f_1}{r_0 f_0},$$

for $r_0, r_1 > 0$. As a result, we may assume without loss of generality that $r_0 = 1$, since our interest lies solely in the ratio $r_1/r_0$, and $r_1 \equiv r \geq 1$; if this condition does not hold, we can simply switch the roles of $f$ and $g$ and consider $1/r$ instead. Now, it is instructive to analyse the behaviour of the sample mean of the permuted data. In this regard, Lemma 7 implies the following unconditional result:

**Corollary 15.** *Let* $N_{Y,1}^\sigma = \sum_{j=n+1}^{n+m} Z_{\sigma(j)}$, *where* $\sigma$ *is sampled according to* (3). *If* $n/m \to \tau > 0$, *then* $\mathbb{E}[(m^{-1} N_{Y,1}^\sigma)^k] \longrightarrow \gamma_1^k$ *for all* $k \in \mathbb{N}$ *as* $n, m \to +\infty$, *where*

$$\gamma_1(f_1, g_1, r, \tau) \equiv \gamma_1 = \begin{cases} \frac{\tau f_1 + g_1}{\tau + 1} + \frac{(r-1)\frac{\tau - 1}{\tau + 1}(\tau f_1 + g_1) + \tau + r - \sqrt{(\tau + r + (r-1)(\tau f_1 + g_1))^2 - 4(r-1)r(\tau f_1 + g_1)}}{2(r-1)} & \text{if } r > 1, \\ \frac{\tau f_1 + g_1}{\tau + 1} & \text{if } r = 1. \end{cases}$$

(55)

*Proof.* Let $\mu$ be the counting measure and equip the set $\{0, 1\}$ with the discrete topology. Lemma 7 in Section 3 gives

$$\frac{1}{m} \sum_{j=n+1}^{n+m} \varphi(Z_{\sigma(j)}) \overset{\mathbb{P}}{\longrightarrow} \int \varphi\{\tau f + g - \tau h_\infty\} d\mu,$$

for every bounded and continuous function $\varphi$ on $\mathcal{X}$, where

$$h_\infty = \frac{\tau f + g}{\tau + \lambda_\infty (r\mathbb{1}\{\cdot = 1\} + \mathbb{1}\{\cdot = 0\})}$$

and $\lambda_\infty > 0$ is the positive solution of

$$1 = \int h_\infty d\mu = \frac{\tau f_1 + g_1}{\tau + \lambda_\infty r} + \frac{\tau f_0 + g_0}{\tau + \lambda_\infty} = \frac{\tau f_1 + g_1}{\tau + \lambda_\infty r} - \frac{\tau f_1 + g_1}{\tau + \lambda_\infty} + \frac{\tau + 1}{\tau + \lambda_\infty}.$$

As $\mathcal{X} = \mathcal{Y} = \{0, 1\}$, we can choose $\varphi = \mathrm{id}$ and obtain that

$$m^{-1} N_{Y,1}^\sigma \xrightarrow{\mathbb{P}} \tau f_1 + g_1 - \tau \frac{\tau f_1 + g_1}{\tau + \lambda_\infty r} = \gamma_1,$$

where in the last step we simply plugged in the expression for $\lambda_\infty$, which can be found explicitly by solving the equation above. We have thus established that $m^{-1} N_{Y,1}^\sigma \xrightarrow{\mathbb{P}} \gamma_1$, which implies the existence of a subsequence along which convergence holds almost surely; applying the dominated convergence theorem yields $\mathbb{E}[(m^{-1} N_{Y,1}^\sigma)^k] \to \gamma_1^k$ for all $k \in \mathbb{N}$ along this subsequence, and by the uniqueness of limits, the convergence extends to the entire sequence. This completes the proof. $\qquad\square$

This result shows that $m^{-1} N_{Y,1}^\sigma \xrightarrow{\mathbb{P}} \gamma_1$, which offers a more explicit interpretation of the result in Lemma 7 and emphasises the intricate dependence of the limiting distribution of the permuted data on the initial parameters $(f_1, g_1, r, \tau)$, even in simple cases. Interestingly, under the null hypothesis, $\gamma_1 = g_1$, as expected. Furthermore, if $r = 1$, we find that $\gamma_1 = (\tau f_1 + g_1)/(\tau + 1)$, reflecting the fact that the DRPT selects permutations uniformly at random when $r = 1$, consistent with (8). Similarly, we can establish an analogous result for the convergence of $n^{-1} \sum_{i=1}^n Z_{\sigma(i)}$ to $\nu_1$, where $\nu_1$ satisfies $\tau f_1 + g_1 = \tau \nu_1 + \gamma_1$, by leveraging the constraint that the total number of ones must remain conserved.

Coming to power results for the case $J \geq 1$, we already know by Theorem 6 that the discrete DRPT is consistent when IPMs which are functions of $(N_{Y,0}^\sigma, \ldots, N_{Y,J}^\sigma)$ are used as test statistics. We now prove another consistency result for a different choice of the test statistic. In this regard, observe that in this setting the null hypothesis (1) is equivalent to

$$H_0 : \frac{g_j}{g_0} = \frac{r_j f_j}{r_0 f_0} \quad \text{for all } j \in [J],$$

for fixed $r = (r_0, r_1, \ldots, r_J) \in R_+^{J+1}$. As before, we can assume $r_0 = 1$ without loss of generality. This motivates the introduction of

$$T(Z_\sigma) = \frac{1}{nm} \sum_{j=1}^J \left| r_j^{-1/2} N_{Y,j}^\sigma (\mathrm{tot}_0 - N_{Y,0}^\sigma) - r_j^{1/2} N_{Y,0}^\sigma (\mathrm{tot}_j - N_{Y,j}^\sigma) \right|, \tag{56}$$

which serves as an estimator of $D(f, g) \equiv D^r(f, g) := \sum_{j \in [J]} \left| r_j^{-1/2} g_j f_0 - r_j^{1/2} f_j g_0 \right|$, which is a population measure of discrepancy that characterises the null. We can prove the following:

**Proposition 16.** *Fix $\alpha \in (0,1)$ and $H > \lceil 1/\alpha - 1 \rceil$. Let $\mathcal{X} = \mathcal{Y} = \{0, \ldots, J\} =: \mathcal{J}$ with $J \geq 1$, $r = (r_0 = 1, r_1, \ldots, r_J) \in R_+^{J+1}$, and let $H_0 : g_j \propto r_j f_j$ for all $j \in \mathcal{J}$. Then, if $n/m \to \tau > 0$, the discrete DRPT using* (56) *as its test statistic is consistent for $H_0$.*

*Proof.* Let $\mu$ be the counting measure, and equip the set $\mathcal{J}$ with the discrete topology. Then, Lemma 7 gives

$$\frac{N_{Y,j}^\sigma}{m} = \frac{1}{m} \sum_{i=n+1}^{n+m} \mathbb{1}\{Z_{\sigma(i)} = j\} \xrightarrow{\mathbb{P}} \int \mathbb{1}\{\cdot = j\}\{\tau f + g - \tau h_\infty\} d\mu$$

$$= \sum_{k=0}^{J} \mathbb{1}\{k = j\} \left\{ \tau f_k + g_k - \tau \frac{\tau f_k + g_k}{\tau + \lambda_\infty r_k} \right\} = \frac{\lambda_\infty r_j}{\tau + \lambda_\infty r_j} (\tau f_j + g_j),$$

where $\lambda_\infty$ is the positive solution of

$$\int \frac{\tau f + g}{\tau + \lambda_\infty r} = \sum_{k=0}^{J} \frac{\tau f_k + g_k}{\tau + \lambda_\infty r_k} = 1.$$

As a result, the generic $j$-th term of the test statistic (56) converges in probability to

$$\frac{1}{nm} \left\{ r_j^{-1/2} N_{Y,j}^\sigma (\text{tot}_0 - N_{Y,0}^\sigma) - r_j^{1/2} N_{Y,0}^\sigma (\text{tot}_j - N_{Y,j}^\sigma) \right\}$$

$$\xrightarrow{\mathbb{P}} r_j^{-1/2} \frac{\lambda_\infty r_j}{\tau + \lambda_\infty r_j} (\tau f_j + g_j) \left\{ \frac{\tau f_0 + g_0}{\tau} - \frac{1}{\tau} \frac{\lambda_\infty r_0}{\tau + \lambda_\infty r_0} (\tau f_0 + g_0) \right\}$$

$$+ r_j^{1/2} \frac{\lambda_\infty r_0}{\tau + \lambda_\infty r_0} (\tau f_0 + g_0) \left\{ \frac{\tau f_j + g_j}{\tau} - \frac{1}{\tau} \frac{\lambda_\infty r_j}{\tau + \lambda_\infty r_j} (\tau f_j + g_j) \right\}$$

$$= r_j^{-1/2} (\tau f_0 + g_0)(\tau f_j + g_j) \frac{\lambda_\infty r_j (1 - r_0)}{(\tau + \lambda_\infty r_0)(\tau + \lambda_\infty r_j)} = 0,$$

since $r_0 = 1$. This shows that $T(Z_\sigma) \xrightarrow{\mathbb{P}} 0$, and since $T(Z) \xrightarrow{\mathbb{P}} D(f, g) > 0$ under the alternative, consistency follows exactly as in the proof of Theorem 6. $\qquad\square$

Finally, coming back again to the case of binary data, the dependence of the minimax separation on $r$ can be analysed more effectively compared to Theorems 9 and 10. In this regard, let $\mathcal{X} = \mathcal{Y} = \{0, 1\}$ and $r \geq 1$, and and consider the measure of discrepancy defined above, i.e. $D(f, g) = \left| r^{-1/2} g_1 f_0 - r^{1/2} f_1 g_0 \right|$. Similarly to Remark 2, this quantity remains unchanged when taking reciprocals — in other words, it is invariant under swapping the zeros and ones, and the $X$'s with the $Y$'s. Now, for fixed $r \geq 1$ and $\rho > 0$, consider

$$H_0 : \frac{g_1}{g_0} = r \frac{f_1}{f_0} \quad \text{vs.} \quad H_1^r(\rho) : D(f, g) > \rho.$$

Write $\Psi$ for the set of all tests, that is randomised functions of $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$, and $\Psi(\alpha)$ for the set of tests of size $\alpha$, with $\alpha \in (0, 1)$. For $\beta \in (0, 1 - \alpha)$, we may define the minimax separation as

$$\rho_r^* \equiv \rho_r^*(n, m, \alpha, \beta) := \inf \left\{ \rho > 0 : \alpha + \inf_{\varphi \in \Psi(\alpha)} \sup_{(f,g) \in H_1^r(\rho)} \mathbb{E}_P(1 - \varphi) \leq \alpha + \beta \right\},$$

where $P = P_f^{\otimes n} \otimes P_g^{\otimes m}$. We now prove a lower bound on $\rho_r^*$.

**Proposition 17.** *Let $\rho_r^*$ the minimax separation defined above and suppose that $\alpha + \beta < 1/2$. We have that $\rho_r^*(n, m, \alpha, \beta) \geq \sqrt{\frac{r}{(n \wedge m)(1+r)^2} \{1 - 2(\alpha + \beta)\}}$.*

*Proof.* For $0 < \rho^2 \leq \frac{r}{(1+r)^2}$, consider

$$
(f^{(0)}, g^{(0)}) = \left( \left( \frac{1}{2}, \frac{1}{2} \right), \left( \frac{1}{1+r}, \frac{r}{1+r} \right) \right) \quad \text{and} \quad (f^{(1)}, g^{(1)}) = \left( \left( \frac{1-\gamma}{2}, \frac{1+\gamma}{2} \right), \left( \frac{1}{1+r}, \frac{r}{1+r} \right) \right),
$$

with $\gamma = \frac{1+r}{\sqrt{r}} \rho$. Observe that $D(f^{(0)}, g^{(0)}) = 0$ and $D(f^{(1)}, g^{(1)}) = \rho$. We will use the well-known fact that the squared Hellinger distance between two discrete probability distributions $p, q$ supported on $[J]$ is given by

$$
\mathrm{H}^2(p, q) = \sum_{i \in [J]} (\sqrt{p_i} - \sqrt{q_i})^2.
$$

Writing $P_{f^{(i)}}$ for the Bernoulli distribution with parameter $f^{(i)}$ for $i \in \{0, 1\}$, we thus have

$$
\mathrm{H}^2(P_{f^{(0)}}, P_{f^{(1)}}) = \frac{1}{2} \left\{ \left( 1 - \sqrt{1-\gamma} \right)^2 + \left( 1 - \sqrt{1+\gamma} \right)^2 \right\} \leq \gamma^2 = \frac{(1+r)^2}{r} \rho^2,
$$

where the last inequality relies on $(1 - \sqrt{1 \pm x})^2 \leq x^2$, and further shows that $\mathrm{H}^2(P_{f^{(0)}}, P_{f^{(1)}}) \leq 1$ since $\rho^2 \leq \frac{r}{(1+r)^2}$. We can then bound the minimax risk using a standard Le-Cam two-point argument as

$$
\alpha + \sup_{(f,g) \in H_1^r(\rho)} \mathbb{E}_P(1 - \varphi) \geq \sup_{(f,g) \in H_0} \mathbb{E}_P \varphi + \sup_{(f,g) \in H_1^r(\rho)} \mathbb{E}_P(1 - \varphi)
$$

$$
\geq 1 - \mathrm{TV}\left( P_{f^{(0)}}^{\otimes n} \otimes P_{g^{(0)}}^{\otimes m}, P_{f^{(1)}}^{\otimes n} \otimes P_{g^{(1)}}^{\otimes m} \right) \geq 1 - \left\{ \mathrm{TV}\left( P_{f^{(0)}}^{\otimes n}, P_{f^{(1)}}^{\otimes n} \right) + \mathrm{TV}\left( P_{g^{(0)}}^{\otimes m}, P_{g^{(1)}}^{\otimes m} \right) \right\}
$$

$$
= 1 - \mathrm{TV}\left( P_{f^{(0)}}^{\otimes n}, P_{f^{(1)}}^{\otimes n} \right) \geq \frac{1}{2} \left( 1 - \frac{1}{2} \mathrm{H}^2(P_{f^{(0)}}, P_{f^{(1)}}) \right)^{2n} \geq \frac{1}{2} \left( 1 - n \, \mathrm{H}^2(P_{f^{(0)}}, P_{f^{(1)}}) \right)
$$

$$
\geq \frac{1}{2} \left( 1 - \frac{n(1+r)^2}{r} \rho^2 \right),
$$

where in the fifth inequality we used the fact that $(1-x)^n \geq 1 - nx$ for $n \in \mathbb{N}$ and $x \leq 1$. The last display is lower bounded by $\alpha + \beta$ if and only if $\rho^2 \leq \frac{r}{n(1+r)^2} \{1 - 2(\alpha + \beta)\}$. Note that $0 < \rho^2 \leq \frac{r}{(1+r)^2}$ is necessarily satisfied since $n \geq 1$ and $0 \leq \alpha + \beta < 1/2$. Switching the roles of $f^{(i)}$ and $g^{(i)}$ in light of the symmetry between the $X$'s and the $Y$'s concludes the proof. $\square$

Proposition 17 suggests that the testing problem is the hardest when $r = 1$. As already mentioned in Section 3, this is accordance with the goodness-of-fit (GoF) testing problem, where the goal is to test the null hypothesis $f = f_0$ for a fixed density $f_0$, based on i.i.d. samples $X_1, \ldots, X_n \sim f$. In the GoF setting, the minimax separation rate depends on the choice of $f_0$, and it has been shown that the problem is hardest when $f_0$ is the uniform distribution (see Balakrishnan and Wasserman, 2019). In complete analogy, and according to the simulation results in Section 5.1, Proposition 17 seems to indicate that $r = 1$ corresponds to the harder testing problem. In other words, more extreme shifts should be easier to detect. We validate this conjecture through simulations on synthetic data. Here, we replicate the setup used in the proof of

Proposition 17, selecting

$$(f_r, g_r) = \left( \left( \frac{1 - \gamma_r}{2}, \frac{1 + \gamma_r}{2} \right), \left( \frac{1}{1+r}, \frac{r}{1+r} \right) \right) \quad \text{with } \gamma_r = \frac{1+r}{\sqrt{r}} \eta.$$

In Figure 8, we assess the performance of the discrete DRPT using (56) as its test statistic over 3000 repetitions with $n = m = 500$, $r \in \{0.1, 0.5, 1, 2, 10\}$ and plotting an estimate of the power function for varying $\eta \in \{0, \dots, 0.1\}$. The results empirically support the tightness of the lower bound of Proposition 17, demonstrating that the power of the discrete DRPT increases when $r$ moves further away from 1.
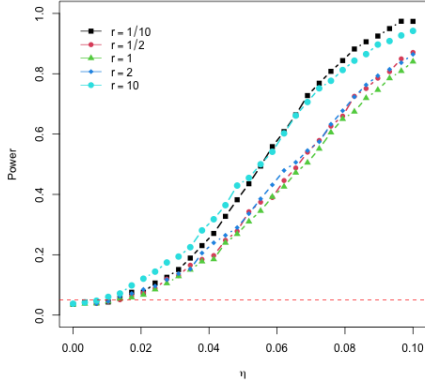


Figure 8: Simulation study with synthetic binary data for varying $r \in \{0.1, 0.5, 1, 2, 10\}$. The discrete DRPT was implemented using (56).
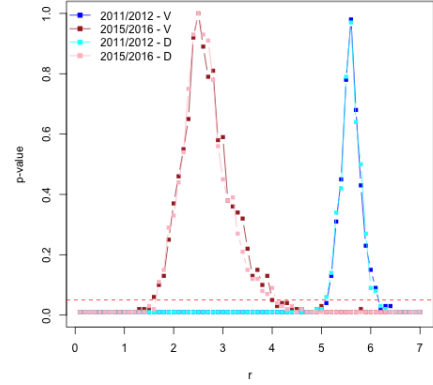


Figure 9: Analogue of Figure 7, also including results for the discrete DRPT based on (56). Legend: V corresponds to the test statistic (11), and D to (56).

Establishing the optimality of the rate $1/\sqrt{r(n \wedge m)}$ is more delicate. While the two-moment method used in the proof of Theorem 9 suffices to derive an upper bound on $\rho_r^*$, it yields a loose dependence on $r$, even though it accurately captures the scaling with $n$ and $m$. Specifically, we can show that $\mathrm{Var}[T(Z)] \lesssim r/n$ and $\mathrm{Var}[T(Z_\sigma)] \lesssim r^2/n$. The first inequality follows from the independence between the $X$'s and the $Y$'s, while the second relies on the following proposition, which we include for completeness.

**Proposition 18.** *Assume $r \geq 1$ and $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. Define $N_{Y,1}^\sigma = \sum_{i=n+1}^{n+m} Z_{\sigma(i)}$, where $\sigma$ is sampled according to (3), and let $\gamma_1$ be as in (55). Then*

$$\mathbb{E}[(N_{Y,1}^\sigma - m\gamma_1)^2] \leq (1+r)n \wedge m + 2f_1(1 - f_1)n + 2g_1(1 - g_1)m.$$

*Proof.* All the expectations are to be intended conditionally to $Z$. We will assume $n = \tau m$ throughout the proof to simplify the computations. Write $\mathcal{T} = \mathrm{tot}_1$, and define

$$p_1^x = \frac{(n - \mathcal{T} + x)x}{(1+r)nm} \quad \text{and} \quad p_2^x = \frac{r(\mathcal{T} - x)(m - x)}{(1+r)nm}.$$
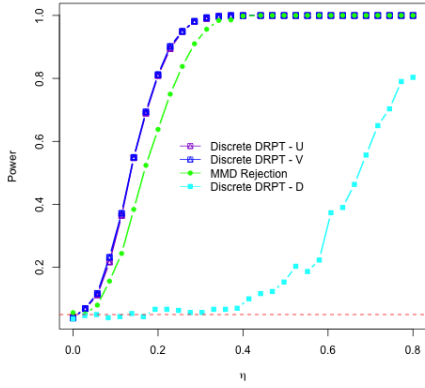
Figure 10: Analogue of Figure 4, also including results for the DRPT based on (56). Legend: U corresponds to the test statistic (12), V to (11), and D to (56).
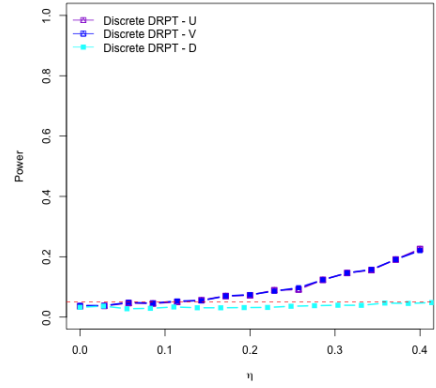


Figure 11: Analogue of Figure 5, also including results for the DRPT based on (56). Legend: U corresponds to the test statistic (12), V to (11), and D to (56).

Let $\gamma(\mathcal{T})$ be the solution to $p_1^x = p_2^x$, i.e.

$$\gamma(\mathcal{T}) = \frac{1}{2(r-1)}\left\{(r-1)\mathcal{T} + n + rm - \sqrt{[(r-1)\mathcal{T} + n - rm]^2 + 4rmn}\right\}.$$

We will show that $\mathbb{E}[(N_{Y,1}^\sigma - \gamma(\mathcal{T}))^2] \leq \frac{1+r}{2}n \wedge m$ and $\mathbb{E}[(m\gamma_1 - \gamma(\mathcal{T}))^2] \leq nf_1(1 - f_1) + mg_1(1 - g_1)$, which imply the desired result since $\mathbb{E}[(N_{Y,1}^\sigma - m\gamma_1)^2] \leq 2\mathbb{E}[\{N_{Y,1}^\sigma - \gamma(\mathcal{T})\}^2] + 2\mathbb{E}[\{m\gamma_1 - \gamma(\mathcal{T})\}^2]$. Now, as for the latter, observe that $m\gamma_1 = \gamma(nf_1 + mg_1)$, and that $\gamma(\cdot)$ is $1-$Lipschitz. This is due to fact that

$$|\gamma'(x)| \leq \frac{1}{2}\left(1 + \frac{|(r-1)x + n - rm|}{\sqrt{[(r-1)x + n - rm]^2 + 4rnm}}\right) \leq 1.$$

Hence,

$$\mathbb{E}[(m\gamma_1 - \gamma(\mathcal{T}))^2] = \mathbb{E}[\{\gamma(nf_1 + mg_1) - \gamma(\mathcal{T})\}^2] \leq \mathbb{E}[\{(nf_1 + mg_1) - \mathcal{T}\}^2]$$

$$= \mathbb{E}\left[\left\{(nf_1 + mg_1) - \sum_{i=1}^n X_i - \sum_{i=1}^m Y_i\right\}^2\right] = \mathrm{Var}\left[\sum_{i=1}^n X_i\right] + \mathrm{Var}\left[\sum_{i=1}^m Y_i\right] = nf_1(1 - f_1) + mg_1(1 - g_1).$$

As for the other term, start by noticing that Algorithm 1 works the same if at every time step $t \in \mathbb{N}$ we just choose a single couple $(i, j)$ with $i \in [n]$ and $j \in \{n + 1, \ldots, n + m\}$ at random, and then switch $Z_{\sigma_t(i)}$ with $Z_{\sigma_t(j)}$ with probability equal to $r^{Z_{\sigma_t(i)}}/(r^{Z_{\sigma_t(i)}} + r^{Z_{\sigma_t(j)}})$. This is not efficient from a computational point of view, but simplifies the proof, since every time step $t$ corresponds at most to one switch. Now, let $K_t$ be the sum of the last $m$ observations after $t$ steps of this simplified algorithm, and observe that it has the same distribution of $N_{Y,1}^\sigma$ for every $t \in \mathbb{N}$ when the procedure is initialised at stationarity. Thus, for $\gamma \equiv \gamma(\mathcal{T})$ to

ease notation, it follows that

$$\mathbb{E}[(K_{t+1} - \gamma)^2 | K_t] - (K_t - \gamma)^2 = p_1^{K_t} + p_2^{K_t} + 2\{(p_2^{K_t} - p_2^\gamma) - (p_1^{K_t} - p_1^\gamma)\}(K_t - \gamma)$$

$$= p_1^{K_t} + p_2^{K_t} - \frac{2(K_t - \gamma)^2}{(1+r)nm}\sqrt{[(r-1)\mathcal{T} + n - rm]^2 + 4rmn} + \frac{2(r-1)}{(1+r)nm}(K_t - \gamma)^3$$

$$\leq 1 - \frac{2(K_t - \gamma)^2}{(1+r)nm}\sqrt{[(r-1)\mathcal{T} + n - rm]^2 + 4rmn} + \frac{2(r-1)(m-\gamma)}{(1+r)nm}(K_t - \gamma)^2$$

$$= 1 - \frac{2(K_t - \gamma)^2}{(1+r)nm}\left\{\sqrt{[(r-1)\mathcal{T} + n - rm]^2 + 4rmn} - (r-1)(m-\gamma)\right\}$$

$$= 1 - \frac{(K_t - \gamma)^2}{(1+r)nm}\left\{\sqrt{[(r-1)\mathcal{T} + n - rm]^2 + 4rmn} + (r-1)\mathcal{T} + n - (r-2)m\right\}$$

$$\leq 1 - \frac{(K_t - \gamma)^2}{(1+r)nm}\left\{|(r-1)\mathcal{T} + n - rm| + (r-1)\mathcal{T} + n - rm + 2m\right\}$$

$$\leq 1 - \frac{2(K_t - \gamma)^2}{(1+r)n} \leq 1 - \frac{2(K_t - \gamma)^2}{(1+r)n}.$$

Taking expectation with respect to $K_t$ under stationarity yields

$$0 \leq 1 - \frac{2}{(1+r)n}\mathbb{E}[\{K_t - \gamma(\mathcal{T})\}^2] = 1 - \frac{2}{(1+r)n}\mathbb{E}[\{N_{Y,1}^\sigma - \gamma(\mathcal{T})\}^2],$$

implying

$$\mathbb{E}[\{N_{Y,1}^\sigma - \gamma(\mathcal{T})\}^2] \leq \frac{1+r}{2}n.$$

By symmetry, we can repeat the same computations for $N_{X,1}^\sigma := \mathcal{T} - N_{Y,1}^\sigma$ and $\nu(\mathcal{T}) := \mathcal{T} - \gamma(\mathcal{T})$ and get

$$\mathbb{E}[\{N_{Y,1}^\sigma - \gamma(\mathcal{T})\}^2] = \mathbb{E}[\{N_{X,1}^\sigma - \nu(\mathcal{T})\}^2] \leq \frac{1+r}{2}m,$$

which gives $\mathbb{E}[\{N_{Y,1}^\sigma - \gamma(\mathcal{T})\}^2] \leq \frac{1+r}{2}n \wedge m$ and concludes the proof. $\qquad\square$

Finally, Figures 9, 10, and 11 present the performance of the discrete DRPT method using the test statistic (56), and compare it with the results from Section 5, where the statistics (11) and (12) were employed. The findings indicate that in the synthetic data scenarios (corresponding to Figures 4 and 5), (56) appears significantly less powerful than the RKHS-based approach discussed in the main text, while in the Frisk example in Figure 9 the two methods seem to be equivalent.