Statistical testing under distributional shifts

Thams, Nikolaj; Saengkyongam, Sorawit; Pfister, Niklas; Peters, Jonas

Published in:
Journal of the Royal Statistical Society, Series B (Statistical Methodology)

DOI:
10.1093/jrsssb/qkad018

Publication date:
2023

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Statistical testing under distributional shifts

Nikolaj Thams, Sorawit Saengkyongam, Niklas Pfister and Jonas Peters

Department of Mathematical Sciences, University of Copenhagen, Copenhagen, Denmark

Address for correspondence: Nikolaj Thams, Universitetsparken 5, 2100 Copenhagen O, Denmark.
Email: thams@math.ku.dk

Abstract

We introduce statistical testing under distributional shifts. We are interested in the hypothesis $P^* \in H_0$ for a target distribution $P^*$, but observe data from a different distribution $Q^*$. We assume that $P^*$ is related to $Q^*$ through a known shift $\tau$ and formally introduce hypothesis testing in this setting. We propose a general testing procedure that first resamples from the observed data to construct an auxiliary data set (similarly to sampling importance resampling) and then applies an existing test in the target domain. We prove that if the size of the resample is of order $o(n^{1/2})$ and the resampling weights are well behaved, this procedure inherits the pointwise asymptotic level and power from the target test. If the map $\tau$ is estimated from data, we maintain the above guarantees under mild conditions on the estimation. Our results extend to finite sample level, uniform asymptotic level, a different resampling scheme, and statistical inference different from testing. Testing under distributional shifts allows us to tackle a diverse set of problems. We argue that it may prove useful in contextual bandit problems and covariate shift, show how it reduces conditional to unconditional independence testing and provide example applications in causal inference.

Keywords: distributional shifts, hypothesis testing, conditional independence testing, sampling importance resampling, causality

1 Introduction

Testing scientific hypotheses about an observed data generating mechanism is an important part of many areas of empirical research and is relevant for almost all types of data. In statistics, the data generating mechanism is described by a distribution $P^*$ and the process of testing a hypothesis corresponds to testing whether $P^*$ belongs to a subclass of distributions $H_0$. In practice, observations from $P^*$, for which we want to test the hypothesis $P^* \in H_0$, may not always be available. For instance, sampling from $P^*$ may be unethical if it corresponds to assigning patients to a certain treatment. $P^*$ may also represent the response to a policy that a government is considering to introduce. Yet, in many cases, one may still have data from a different, but related, distribution $Q^*$. In the examples above, this could be data from an observational study or under the policies currently deployed by the government.

Although specialised solutions exist for many such problems, there is no general method for tackling them. In this paper, we aim to analyse the above testing task from a general point of view. We assume that a distributional shift $\tau : Q \to P$ between domains $Q$ and $P$ is known and, using data from $Q^*$, aim to test the hypothesis $P^* = \tau(Q^*) \in H_0$. We propose the following general framework. We resample from $Q^*$ to construct an auxiliary data set mimicking a sample from $P^*$ and then exploit the existence of a test in the target domain. We assume that $Q^*$ and $P^*$ are absolutely continuous with respect to the same background product measure. Our method does not assume full knowledge of $Q^*$ or $P^*$, but only knowledge of the (potentially unnormalised) ratio

---

Received: April 28, 2022. Revised: August 30, 2022. Accepted: November 10, 2022


This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
\( p^*/q^* \), where \( q^* \) and \( p^* \) are densities of \( Q^* \) and \( P^* \) with respect to the background measure, respectively. If, for example, the shift corresponds to a change in the conditional distribution of a few of the observed variables, one only needs to know these changing conditionals.

Our framework assumes the existence of a test \( \varphi \) in the target domain, i.e., a test that could be applied if data from \( P^* \) were available. This test \( \varphi \) is then applied to a resampled version of the observed data set. Here, we propose a sampling scheme that is similar to sampling importance resampling (SIR), proposed by Rubin (1987) and Smith and Gelfand (1992) but generates a distinct sample of size \( m \), using weights \( r(X) \propto q^*(X)/p^*(X) \) where \( X \) is a random vector with distribution \( Q^* \). We prove that this procedure inherits the pointwise asymptotic properties of the test \( \varphi \) if the weights \( r \) have finite second moment in \( Q^* \), and \( m = o(\sqrt{n}) \). In particular, the procedure holds pointwise asymptotic level if the test \( \varphi \) does. We show that the same can be obtained if \( r \) is not known, but can be estimated from data sufficiently well. The proposed method is easy-to-use and can be applied to any hypothesis test, even if the test is based on a nonlinear test statistic.

Several problems can be cast as hypothesis tests under distributional shifts. This includes hypothesis tests in off-policy evaluation, tests of conditional independence, testing the absence of causal edges through dormant independences (Shpitser & Pearl, 2008; Verma & Pearl, 1991), that is, testing certain equality constraints in an observed distribution, and problems of covariate shift. Our proposed method can be applied to all of these problems. For some of them, we are not aware of other methods with theoretical guarantees—this includes dormant independence testing with continuous variables, off-policy testing with complex hypotheses, conditional independence testing with censored data and model selection under covariate shift with complex scoring functions. The framework also inspired a novel method for causal discovery that exploits knowledge of a single causal conditional. For some of the above problems, however, more specialised solutions exist, and as such, the proposed testing procedure relates to a line of related work.

Ratios of densities have been applied in the reinforcement learning literature (e.g., Sutton & Barto, 1998), where inference in \( P^* \) using data from \( Q^* \) is known as off-policy prediction (Precup et al., 2001). One can estimate the expectation of \( X \) under \( P^* \) using importance sampling, (IS), that is, as weighted averages using weights \( r(X) \) possibly truncated to decrease the variance of the estimation (Mahmood et al., 2014; Precup et al., 2001). An approach to estimate means in \( P^* \) based on resampling was proposed by Schlegel et al. (2019). However, they consider the size of the resample fixed and do not consider statistical testing. Thomas et al. (2015) propose bootstrap confidence intervals for off-policy prediction based on important weighted returns. Hao et al. (2021) present a bootstrapping approach with Fitted Q-Evaluation (FQE) for off-policy statistical inference and demonstrate its distributional consistency guarantee.

In the causal inference literature, inverse probability weighting (IPW) can be used to adjust for confounding or selection bias in data (e.g., Horvitz & Thompson, 1952; Robins et al., 2000). To estimate the effect of a treatment \( X \) on a response \( Y \), one can weight each observed response \( Y_i \) with \( 1/q^*(X_i|Z_i) \), where \( Z_i \) is an observed confounder. For continuous treatments, it has been proposed to change the numerator to a marginal distribution \( p^*(x) \) to stabilise the weights (Hernán & Robins, 2006; Naimi et al., 2014). Both choices of weights appear in our framework, too (e.g., the first one corresponds to a target distribution with \( p^*(x|z) \propto 1 \)). In general, IS and IPW can only be applied if the population version of the test statistic can be written as a mean of a function of a single observation, such as \( E[Y_i] \) or \( E[f(X_i, Z_i)] \), whereas our approach also applies to test statistics that are functions of the entire sample, which is the case for many tests that go beyond testing moments, such as several independence tests, for example.

SIR sampling schemes were first studied by Rubin (1987) and are often used in the context of Bayesian inference (Smith & Gelfand, 1992). Skare et al. (2003) show that when using weighted resampling with or without replacement, for \( n \to \infty \) and fixed \( m \), the sample converges towards \( m \) i.i.d. draws from the target distribution, and provide rates for the convergence. Our work is inspired by these types of results, even though our proofs require different techniques.

Our paper adds to the literature on distributional shifts by considering hypothesis tests in shifted distributions. In the context of prediction, distributional shifts, or dataset shifts, have been studied in the machine learning literature both to handle the situations where a marginal covariate distribution changes and when the conditional distribution of label given covariate changes (Quiñonero-Candela et al., 2009). If the shift represents a changing marginal distribution and unlabelled samples are available from both training and test environments, Huang et al. (2006)
proposed kernel mean matching, which nonparametrically reweights the training loss to resemble the loss on a target sample. In settings where a generative model and causal graph is known, Pearl and Bareinboim (2011) and Subbaswamy et al. (2019) provide graphical criteria under which causal estimates can be ‘transported’ from one distribution $Q$ to a shifted distribution $P$, assuming knowledge of both joint distributions $Q$ and $P$. In contrast, we consider statistical testing, and neither assume knowledge of the full causal graph nor availability of samples from the target distribution, but instead knowledge of how the target data differ from the observed data. Conformal prediction (Vovk et al., 2005) has been applied to covariate shift (Park et al., 2021; Tibshirani et al., 2019) too, but its goal of constructing a prediction interval for a new random observation is generally different than the one of the proposed testing framework (see, however, the extension of our theory to general inference problems, Section 4.6).

In this paper, we first formally define testing under distributional shifts, and define notions such as pointwise and asymptotic level when using observed data to test the hypothesis in the target domain. Second, we outline a number of statistical problems, that can be solved by testing under a shift, including conditional independence testing and testing dormant independences. Third, we propose methods that enable testing under distributional shifts: both a simple method based on rejection sampling and a resampling scheme that requires fewer assumptions than the rejection sampler. Fourth, we provide finite sample and asymptotic guarantees for our proposed resampling scheme; contrary to the existing literature, where typically $m$ fixed and $n \to \infty$ has been studied (e.g., Skare et al., 2003), we study the asymptotic behaviour of our resampling test when both $m$ and $n$ approach infinity, and show that under any resampling scheme, the requirement $m = o(\sqrt{n})$ is necessary.

The remainder of this work is structured as follows. We formally introduce the framework of testing under distributional shifts in Section 2. Section 3 showcases how several problems from different statistical fields fit into this framework. We introduce a general procedure for testing under distributional shifts and provide theoretical results in Section 4. In Section 5, we conduct simulation experiments and Section 6 concludes and discusses future work.

## 2 Statistical testing under distributional shifts

### 2.1 Testing hypotheses in a target distribution

Consider a set of distributions $\mathcal{P}$ on a target domain $\mathcal{X} \subseteq \mathbb{R}^d$ and a null hypothesis $H_0 \subseteq \mathcal{P}$. In hypothesis testing, we are usually given data from a distribution $P^* \in \mathcal{P}$ and want to test whether $P^* \in H_0$. In this paper, we consider the problem of testing the same hypothesis but instead of observing data from $P^*$ directly, we assume the data are generated by a different, but related, distribution $Q^*$ from a set of distributions $Q$ over a (potentially) different observational domain $\mathcal{X} \subseteq \mathbb{R}^e$.

More formally, we assume that we have observed data $X_n := (X_1, \ldots, X_n) \in \mathcal{X}^n$ consisting of $n$ i.i.d. random variables $X_i$ with distribution $Q^* \in Q$. We use superscripts to denote the individual coordinates of $X = (X^1, \ldots, X^d) \in \mathcal{X}$. We assume that $Q^*$ and $P^*$ are related through a map $\tau : Q \to \mathcal{P}$, called a (distributional) shift, which satisfies $P^* = \tau(Q^*)$. We aim to construct a randomised hypothesis test $\psi_n : \mathcal{X}^n \times \mathbb{R} \to \{0, 1\}$ that we can apply to the observed data $X_n$ to test the null hypothesis

$$\tau(Q^*) \in H_0.$$  

We reject this null hypothesis if $\psi_n = 1$ and do not reject the null if $\psi_n = 0$. To allow for random components, we let $\psi_n$ take as input a uniformly distributed random variable $U$ (assumed to be independent of the other variables) that generates the randomness of $\psi_n$. Whenever there is no ambiguity about the randomisation, we omit $U$ and write $\psi_n(X_n)$; unless stated otherwise, any expectation or probability includes the randomness of $U$. For $\alpha \in (0, 1)$, we say that $\psi_n$ holds level $\alpha$ at sample size $n$ if it holds that

$$\sup_{Q \in \tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n(X_n, U) = 1) \leq \alpha.$$  


In practice, requiring level at sample size $n$ is often too restrictive. We say that the test has point-wise asymptotic level $\alpha$ if

$$\sup_{Q \in \tau^{-1}(H_0)} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n(X_n, U) = 1) \leq \alpha.$$  

We illustrate the setup in Figure 1.

**Remark.** The map $\tau: Q \to P$ above represents a view that starts with the distribution $Q^*$ of the observed data and considers the distribution $P^*$ of interest as the image under $\tau$. Alternatively, one may also start with a map $\eta: P \to Q$ and say that the test holds level $\alpha$ at sample size $n$ if

$$\sup_{P \in H_0} \mathbb{P}_{\eta(P)}(\psi_n(X_n, U) = 1) \leq \alpha.$$  

This corresponds to a level guarantee for a test of the hypothesis $\eta^{-1}(Q^*) \cap H_0 \neq \emptyset$. If $\tau$ is invertible, the two views trivially coincide with $\eta = \tau^{-1}$, but in general there are subtle differences, see Appendix A.1 for details. In this paper, we use the formulation based on $\tau: Q \to P$, that is, Equations (1), (2) and (3).

### 2.2 Distributional shifts

We consider two types of maps $\tau: Q \to P$, both of which can be written in product form. First, assume that there is a subset $A \subseteq \{1, \ldots, d\}$ together with a known map $r: x^A \mapsto r(x^A) \in [0, \infty)$ such that for all $q \in Q$ the target density $\tau(q)$ satisfies that

$$\tau(q)(x^1, \ldots, x^d) \propto r(x^A) \cdot q(x^1, \ldots, x^d) \quad \text{for all } (x^1, \ldots, x^d) \in \mathbb{Z}.$$  

Second, assume that there is a subset $A \subseteq \{1, \ldots, d\}$ together with a known map $r_q: (q, x^A) \mapsto r_q(x^A) \in [0, \infty)$ such that for all $q \in Q$, the target density $r(q)$ satisfies that

$$\tau(q)(x^1, \ldots, x^d) \propto r_q(x^A) \cdot q(x^1, \ldots, x^d) \quad \text{for all } (x^1, \ldots, x^d) \in \mathbb{Z}.$$  

In the remainder of this work, we assume that $X$ and $Z$ are both subsets of $\mathbb{R}^d$, that is $e = d$, and that all distributions in $P$ and $Q$ have densities with respect to the same dominating product measure $\mu$. We refer to a distribution $Q$ and its density $q$ interchangeably.
Here, we assume that the factor \( r(\cdot) \) can be evaluated for any given \((q, x^k)\). This case arises, for example, when the training policy or the conditional is unknown and needs to be estimated from data. Section 3 contains examples for both types of shifts. If, in any of the above two cases, the set \( A \) is not mentioned explicitly, we implicitly assume \( A = \{1, \ldots, d\} \). In many applications \( r \) represents a local change in the system, so even though \( d \) may be large, \(|A|\) will be much smaller than \( d \). In particular, we do not need to know the entire distribution to evaluate \( r(x^k) \).

In principle, this approach applies to any full-support distribution \( Q \), since for a given target distribution \( P \in \mathcal{P} \), \( P = \pi(Q) \) is satisfied as long as we define \( r(x) = p(x)/q(x) \), and in the case that we consider a change in a single conditional, this simplifies to \( r(x_{d_1:a_2}) = p(x_{d_1:a_2})/q(x_{d_1:a_2}) \). In practice, there may be regions of the support where \( q(x) \) is much smaller than \( p(x) \), in which case the weights will be ill-behaved. We address this issue in (A2) and analyse its impact in Theorem 4. For some shifts and hypotheses, direct solutions are available that do not use the importance weights (5): For example, when testing a hypothesis about \( X^1 \) under a mean shift in the marginal distribution of \( X^1 \), one could directly add the anticipated shift in mean to every observation before testing. However, in most cases involving shifts in conditional distributions or in variables different from those entering the test, such approaches fail. If \( r \) is misspecified, in the sense that \( \pi(Q^*) \neq P^* \), then the guarantees for the methodology below still hold, but for testing the distribution \( \pi(Q^*) \in H_0 \) instead of \( P^* \in H_0 \).

2.3 Exploiting a test in the target domain

In this work, we assume that there is a test \( \varphi \) for the hypothesis \( H_0 \) that can be applied to data from the target domain \( \mathcal{Z} \). Formally, we consider a sequence \( \varphi_k : \mathbb{Z}_k \times \mathbb{R} \rightarrow \mathbb{R} \) of (potentially randomised) hypothesis tests for \( H_0 \) that can be applied to \( k \) observations \( Z_k \) from the target domain \( \mathcal{Z} \) and a uniformly distributed random variable \( V \), generating the randomness of \( \varphi_k \). For simplicity, we omit \( V \) from the notation and write \( \varphi_k(Z_k) \). We say that \( \varphi := (\varphi_k)_k \) has pointwise asymptotic level \( \alpha \) for \( H_0 \) in the target domain if

\[
\sup_{r \in H_0} \limsup_{k \rightarrow \infty} \mathbb{P}_P(\varphi_k(Z_k) = 1) \leq \alpha. \tag{7}
\]

To address the problem of testing under distributional shifts, we propose in Section 4 to resample a data set of size \( m \) from the observed data \( X_m \) (using resampling weights that depend on the shift) and apply the test \( \varphi_m \) to the resampled data. We show that this yields a randomised test \( \psi \), which inherits the pointwise asymptotic properties of \( \varphi_m \) and in particular satisfies the level requirement (3) if \( \varphi \) has pointwise asymptotic level. This procedure is easy-to-use and can be combined with any testing procedure \( \varphi \) from the target domain.

2.4 Testing hypotheses in the observed domain

The framework of testing hypotheses in the target distribution can be helpful even if we are interested in testing a hypothesis about the observed distribution \( Q^* \), that is, testing \( Q^* \in H_0^Q \) for some \( H_0^Q \subseteq Q \). If \( \pi(H_0^Q) \subseteq H_0^P := H_0 \), any test \( \psi_n \) satisfying pointwise asymptotic level (3) for \( H_0^P \subseteq \mathcal{P} \) can be used as a test for \( Q^* \in H_0^Q \), and will still satisfy asymptotic level, see Section 4.4.5.

Such an approach can be particularly interesting when it is more difficult to test \( Q^* \in H_0^Q \) in the observed domain than it is to test \( \pi(Q^*) \in H_0^P \) in the target domain. For example, testing conditional independence in the observed domain can be reduced to (unconditional) independence testing in the target domain. Here, we may benefit from transferring the test into the target domain if one of the conditionals is known or can be estimated from data. Also testing a Verma equality (Verma & Pearl, 1991) in the observed distribution can be turned into an independence test in the target distribution, too; but here, testing directly in the observed domain may not even be possible. Often there is a computational advantage of our approach: In many situations, the resampled data set, where the hypothesis is easier to test, is much smaller than the original data set, see for instance the experiment in Section 5.4. When the hypothesis of interest is in the observed domain, usually different choices for the target distributions are possible. In practice, it is helpful to choose a target distribution that yields well-behaved resampling weights (see (10)), which can often be achieved by
matching certain marginals, see, e.g., Section 5.6 (see also Hernán & Robins, 2006; Robins et al., 2000).

The following Section 3 discusses the above and other applications of testing under distributional shifts in more detail. Corresponding simulation experiments are presented in Section 5. Section 4 provides details of our method and its theoretical guarantees.

3 Example applications of testing under distributional shifts

3.1 Conditional independence testing

Let us first consider a random vector \((X, Y, Z)\) with joint probability density function \(q^*\) and assume that the conditional \(q^*(z|x)\) is known. We can then apply our framework to test

\[
H_0^Q = \{Q : X \perp Y \mid Z \text{ and } q(z|x) = q^*(z|x)\} \quad \text{(cond. ind. in observed domain)}
\]

by reducing the problem to an unconditional independence test. The key idea is to factor a density \(q \in H_0^Q\) as \(q(x, y, z) = q(y|x, z)q^*(z|x)q(x)\), replace\(^3\) the conditional \(q^*(z|x)\) by, e.g., a standard normal density \(\phi(z)\) to obtain the target density \(p\), and then test for unconditional independence of \(X\) and \(Y\). When \(X\) is a randomised treatment, \(Y\) the outcome, and \(Z\) a mediator, this corresponds to testing (nonparametrically) the existence of a direct causal effect (e.g., Hernán & Robins, 2020; Imbens & Rubin, 2015; Pearl, 2009).

Formally, we define a corresponding hypothesis in the target domain:

\[
H_0^P := \{P : X \perp Y \mid p(z|x) = \phi(z)\} \quad \text{(ind. int. target domain)}
\]

with \(\phi\) being the standard normal density. We can then define a map \(\tau\) by

\[
\tau(q)(x, y, z) := \frac{\phi(z)}{q^*(z|x)} \cdot q(x, y, z) \quad \text{for all } (x, y, z) \in Z.
\]

Considering any \(q \in H_0^Q\) and writing \(p := \tau(q)\), we have

\[
p(x, y, z) = \frac{\phi(z)}{q^*(z|x)} q(y|x, z)q^*(z|x)q(x) = q(y|x, z)q(x)\phi(z).
\]

This shows\(^4\) that conditional independence \(X \perp Y \mid Z\) in \(Q\) implies independence \(X \perp Y\) in \(P\) and therefore \(\tau(H_0^Q) \subseteq H_0^P\). Starting with an independence test \(\varphi_m\) for \(H_0^Q\), we can thus test \(\tau(Q^*) \in H_0^P\), with level guarantee in (3). As we have argued in Section 2.4, this corresponds to testing \(Q^* \in H_0^Q\), and thereby reduces the question of conditional independence to independence.

If, instead of \(q^*(z|x)\), we know the reverse conditional \(q^*(x|z)\), we can use the same reasoning as above using the factorisation \(q(x, y, z) = q(z)q^*(x|z)q(y|x, z)\) and a marginal target density \(\phi(x)\) to again test \(X \perp Y \mid Z\). When \(X\) is a treatment, \(Y\) the outcome, \(Z\) the full set of covariates, and \(q^*(x|z)\) represents the randomisation scheme, this corresponds to testing (nonparametrically) the existence of a total causal effect (e.g., Peters et al., 2017) between \(X\) and \(Y\).

If neither of the conditionals is known, we can still fit the test into our framework. To do so, define the hypotheses \(H_0^Q := \{Q : X \perp Y \mid Z\}\), \(H_0^P := \{P : X \perp Y \mid p(z|x) = \phi(z)\}\), and the map \(\tau\) via \(\tau(q)(x, y, z) := \frac{\phi(z)}{q(z|x)} \cdot q(x, y, z)\), for all \((x^1, \ldots, x^d) \in Z\); cf. (6). Section 4.2 shows that one can estimate the conditional \(q(z|x)\) from data and may still maintain the level guarantee of the overall procedure. There are other, more specialised conditional independence tests but this viewpoint may be an interesting alternative if we can estimate one of the conditionals well, e.g., because there

\[^3\] If the factorisation happens to correspond to the factorisation using a causal graph, this is similar to performing an intervention on \(Z\), see Appendix A.2. However, the proposed factorisation is always valid, so this procedure does not make any assumptions about causal structures.

\[^4\] The following statement holds because, clearly, \(p(z|x) = \phi(z)\) and if \(X \perp Y \mid Z\) in \(q\), that is, \(q(y|x, z) = g(y|z)\) for all \(x, y, z\) yielding this expression well defined, it follows \(p(x, y) = p(x)p(y)\).
are many more observations of \((X, Z)\) than there are of \((X, Z, Y)\). Furthermore, conditional independence tests may not even be available in some complex settings, while marginal independence tests may exist. For example, we illustrate in Section 5.5 that our method can be applied to conduct a nonparametric test for conditional independence with right-censored data. To the best of our knowledge, there are no other nonparametric conditional independence tests available in this setting.

The assumption of knowing one conditional is also exploited by the conditional randomisation test (CRT) and the conditional permutation test (CPT) by Candès et al. (2018) and Berrett et al. (2020), respectively. These methods, however, require knowledge of \(q^*(x|z)\) and cannot exploit knowledge of \(q^*(z|x)\). They simulate (in case of CRT) or permute (in case of CPT) \(X\) while keeping \(Z\) and \(Y\) fixed and construct \(p\)-values for the hypothesis of conditional independence. The approaches are similar in that they use the known conditional to create weights. Our method, however, explicitly constructs a target distribution and, as argued above, cannot only exploit knowledge of \(q^*(x|z)\) but also knowledge of \(q^*(z|x)\).

Our approach can be modified to obtain a double robustness property. Suppose we consider a map \(\tau(q)(x, y, z) = q(x, y, z) \cdot \frac{\phi_X(x) \phi_Y(y)}{q(x|z)q(y|z)}\), for some density functions \(\phi_X(x)\) and \(\phi_Y(y)\) and model conditionals \(q(x|z)\) and \(q(y|z)\). Then

\[
\tau(q^*)(x, y, z) = q^*(z)q^*(x, y|z) \cdot \frac{\phi_X(x)\phi_Y(y)}{q(x|z)q(y|z)} = q^*(z)\phi_X(x)\phi_Y(y) \cdot \frac{q^*(x, y|z)}{q(x|z)q(y|z)}.
\]

Suppose, for example, that \(q(x|z) = q^*(x|z)\), but \(q(y|z) \neq q^*(y|z)\). Then, it still holds under the null hypothesis \(X \perp Y|Z\) in \(Q^*\) that \(\tau(q^*)(x, y, z) = \phi_X(x)q^*(z)\phi_Y(y)q^*(y|z)\), meaning that \(X \perp Y\) in \(\tau(Q^*)\). Similarly, if \(q(y|z) = q^*(y|z)\), but \(q(x|z) \neq q^*(x|z)\), conditional independence in \(Q^*\) also implies independence in \(\tau(Q^*)\). That is, as long as one of the modelled conditionals \(q(x|z)\) and \(q(y|z)\) equals the corresponding one of \(q^*\) (we do not need to know which one), the hypothesis of conditional independence in \(Q^*\) can be tested as a hypothesis of marginal independence in \(P^*\). This is similar to the doubly robustness guarantee in Shi et al. (2021), where as long as one estimates at least one conditional consistently, the overall test is consistent.

### 3.2 Off-policy testing

Consider a contextual bandit setup (e.g., Agarwal et al., 2014; Langford & Zhang, 2008). In each round, an agent observes a context \(Z := (Z^1, \ldots, Z^d)\) and selects an action \(A \in \{a_1, \ldots, a_L\}\), based on a known policy \(q^*(a|z)\). The agent then receives a reward \(R\) depending on the chosen action \(A\) and the observed context \(Z\). Suppose we have access to a data set \(X_n\) of \(n\) rounds containing observations \(X_i := (Z_i, A_i, R_i), i = 1, \ldots, n\). We can then test statements about the distribution under another policy \(p^*(a|z)\). For example, we can test whether the expected reward is smaller than zero. To do so, we define

\[
H_0 := \{p : E_p[R] \leq 0 \text{ and } p(a|z) = p^*(a|z)\}
\]

and \(\tau(q)(x) := r(x)q(x)\) with the shift factor \(r(z, a) := p^*(a|z)/q^*(a|z)\). Here, the function of interest can be written as an expectation of a single observation, so other, simpler approaches such as IS or IPW can be used, too (see Section 1).

But it is also possible to test more involved hypotheses. This includes testing (conditional) independence under a new policy, for example. Suppose that one of the covariates \(Z^i\) is used for selecting actions by an observed policy \(q^*(a|z)\). This creates a dependence between \(Z^i\) and \(R\), but it is unclear whether this dependence is only due to the action \(A\) being based on \(Z^i\), or whether \(Z^i\) also depends on \(R\) in other ways, for instance in that \(Z^i\) has a direct effect on \(R\). To test the latter statement, we can create a new policy \(p^*(a|z)\) that does not use \(Z^i\) for selecting actions. Then, we can test whether, under \(p^*(a|z)\), \(R\) is independent of \(Z^i\), given the other variables that the action is based on. If not, we know that there must be a dependence between \(R\) and \(Z^i\) under \(q^*(a|z)\) beyond the action \(A\) being based on \(Z^i\). This may be relevant for learning sets of features that are invariant across different environments, that is, features \(Z^i\) such that \(R \perp Z^i\) is stable across environments.
A policy that depends on such invariant features is guaranteed to generalise to unseen environments (Saengkyongam et al., 2021). Another, more involved hypothesis for off-policy evaluation compares the reward distributions under two different policies. This can be written as a two-sample test, which we discuss in Section 3.3.

This procedure extends to more general reinforcement learning settings, where for example one repeatedly observes a Markov decision process (Sutton & Barto, 1998). The weights then correspond to a product containing one factor for each decision. If the decision process contains many decisions or the data generating policy is not sufficiently close to the policy to be evaluated, off-policy evaluation becomes a difficult problem, with weights not being well behaved. This problem is well known in the reinforcement learning literature (Levine et al., 2020; Mahmood et al., 2014), where the ill-behaved weights result in large variance of estimators; in the methodology we propose in Section 4, it generally results in a loss of power.

3.3 Two-sample testing with one transformed sample

We can use the framework to perform a two-sample test, after transforming one of the two samples. Consider the observed distribution $q^*$ over $X = (X^1, \ldots, X^d) \in \mathbb{R}^d$ and $K \in \{1, 2\}$, where the latter indicates which of the two samples a data point belongs to. We now keep the first sample as it is and change the second sample with a transformation $\tau$, $q^* \mapsto \tau(q^*)$. We can then test whether, after the transformation, the two samples come from the same distribution, i.e., whether

$$q^*(x|k = 1) = \tau(q^*)(x|k = 2)$$

for all $x$. For example, assume that we know the conditional $q^*(x^1|x^2, k = 2)$ and consider transforming this to $p^*(x^1|x^2, k = 2)$. To formally apply our framework, we then define

$$H_0 := \{P : (X^1, \ldots, X^d)_{|K=1} \overset{\mathcal{L}}{=} (X^1, \ldots, X^d)_{|K=2}, \quad p(x^1|x^2, k = 2) = p^*(x^1|x^2, k = 2)\}$$

and the shift $\tau(q)(x^1, \ldots, x^d, k) := r(x^1, x^2, k) \cdot q(x^1, \ldots, x^d, k)$, where

$$r(x^1, x^2, k) := \begin{cases} \frac{1}{q^*(x^1|x^2, k = 2)} & \text{if } k = 1 \\ \frac{1}{q^*(x^1|x^2, k = 2)} & \text{if } k = 2. \end{cases}$$

In particular, this approach can be used for off-policy evaluation (the setting described in the previous section) to test whether the reward under the training policy $q^*(a|z)$ has the same distribution as under a target policy $p^*(a|z)$. We first randomly split the training sample into two subsamples ($K = 1$ and $K = 2$) and then test whether the distribution of the reward is different under the two policies,

$$H_0 := \{P : R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2}, p(a|z, k = 1) = q^*(a|z) \text{ and } p(a|z, k = 2) = p^*(a|z)\},$$

by using weights $r(a, z, k) = p^*(a|z)/q^*(a|z)$ when $k = 2$ and $r(a, z, k) = 1$ otherwise. This is not confined to testing identical distributions: For example, we can also test, nonparametrically, whether the expected reward under the new policy $p^*(a|z)$ is larger than under the current policy $q^*(a|z)$. To do so, we define

$$H_0 := \{P : \mathbb{E}_P[R_{|K=2}] \leq \mathbb{E}_P[R_{|K=1}], p(a|z, k = 1) = q^*(a|z) \text{ and } p(a|z, k = 2) = p^*(a|z)\}.$$ 

Section 5.3 shows some empirical evaluations of such tests.

3.4 Dormant independences

Let us consider a random vector $(X^1, \ldots, X^d)$ with a distribution $Q$ that is Markovian with respect to a directed acyclic graph and that has a density w.r.t. a product measure. By the global
Markov condition (e.g., Lauritzen, 1996), we then have for all disjoint subsets \(A, B, C \subset \{1, \ldots, d\}\) that \(X^A \perp X^B \mid X^C\) if \(A\) \((-\text{separates})\) \(B\) given \(C\). If some of the components of the random vector are unobserved, the Markov assumption still implies conditional independence statements in the observational distribution. In addition, however, it may impose constraints on the observational distribution that are different from conditional independence constraints. Figure 2 shows a famous example, due to Verma and Pearl (1991), that gives rise to the Verma-constraint: If the random vector \((X^1, X^2, X^3, X^4, H)\) has a distribution \(Q\) that is Markovian w.r.t. the graph \(\mathcal{G}\) shown in Figure 2 (left), there exists a function \(f\) such that, for all \(x^1, x^3, x^4\),

\[
\int_{-\infty}^{\infty} q(x^2 | x^1) q(x^4 | x^1, x^2, x^3) \, dx^2 = f(x^3, x^4) \tag{8}
\]

(in particular, \(f\) does not depend on \(x^1\)). This constraint cannot be written as a conditional independence constraint in the observational distribution \(Q\). In general, the constraint (8) does not hold if \(Q\) is Markovian w.r.t. \(\mathcal{H}\) (see Figure 2, right). Assume now that the conditional \(q(x^3 | x^2) = q^{*}(x^3 | x^2)\) is known (e.g., through a randomisation experiment). We can then hope to test for this constraint by considering the null hypothesis

\[
H^0_Q := \{Q : Q \text{satisfies}(8) \text{ and } q(x^3 | x^2) = q^{*}(x^3 | x^2)\}
\]

and hence distinguish between \(\mathcal{G}\) and \(\mathcal{H}\). Constraints of the form (8) have been studied recently, and in some cases the constraints can be exploited to construct score-based structure learning methodology (Nowzohour et al., 2017; Shpitser et al., 2012). Robins (1999) reformulates (8) into a constraint on the conditional expectation of a given test function, and in particular that this conditional expectation does not depend on \(X^4\); to test this in practice, Robins (1999) proposes to use structured nested models. Shpitser and Pearl (2008) show that some of such constraints, called dormant independence constraints, can be written as a conditional independence constraint in an interventional distribution (see also Richardson et al., 2017; Robins, 1999; Shpitser et al., 2014), and Shpitser et al. (2009) propose an algorithm that detects constraints that arise due to dormant independences using oracle knowledge. The Verma constraint (8), too, is a dormant independence, that is, we have

\[
X^1 \perp X^4 \text{ in } Q^{do(X^3 := N)}, \tag{9}
\]

where \(N \sim \mathcal{N}(0, 1)\), for example. Here, \(Q^{do(X^3 := N)}\), denotes the distribution in which \(q^{*}(x^3 | x^2)\) is replaced by \(\phi(x^3)\) see Appendix A.2 for details. Using the described framework, we can test (9) to distinguish between \(\mathcal{G}\) and \(\mathcal{H}\).

In practice, we may need to estimate the corresponding conditional, such as \(q(x^3 | x^2)\) in the example above, from data; as before, this still fits into the framework using (6), see Section 5.6 for a simulation study. In special cases, such as binary, applying resampling methodology to this type of problem has been considered before (Bhattacharya, 2019).

The problem of testing (conditional) independences under an interventional distribution has been shown to be relevant in real-world applications. One application is testing direct effects in

\[\text{...}\]

\[\text{...}\]
dynamic treatment regimes (e.g., Robins & Wasserman, 1997). Consider a two-stage dynamic treatment regime consisting of a sequence of two treatment variables $A^0, A^1$, an intermediate covariate vector $Z^1$ (corresponding to $X^1, X^2$, and $X^1$ in Figure 2, respectively), and an outcome variable $Y$. To test whether there is a direct causal effect of the first treatment $A^0$ on the outcome $Y$, one needs to test whether $A^0$ is independent of $Y$ in the interventional distribution in which the conditional $q^*(z^1|a^1)$ is replaced by some marginal $\phi(z^1)$. Here, one cannot disentangle the direct causal effect of $A^1$ on $Y$ and the total causal effect based solely on the (conditional) independence constraints in the observational distribution; the problem is related to the g-null paradox (McGrath et al., 2022; Robins & Wasserman, 1997).

Another application is testing front-door assumptions in causal effect estimation, which has been proposed by Bhattacharya and Nabi (2022). The front-door adjustment (Pearl, 2009) is an adjustment strategy for identifying the causal effect of a treatment $X$ on an outcome $Y$ in the presence of hidden confounders $U$. The main assumption is that we observe an intermediate variable $M$ that mediates the effect of $X$ on $Y$ and is not confounded by $U$. Bhattacharya and Nabi (2022) consider the setup in which we additionally observe an anchor variable $Z$ which causes both the treatment $X$ and the mediator $M$, but does not directly cause the outcome $Y$. The anchor variable allows for testing of the front-door assumptions by testing independence constraints under the interventional distribution in which the conditional $q^*(m|x, z)$ is replaced by some marginal $\phi(m)$. We follow Bhattacharya and Nabi (2022) and conduct a numerical experiment on the Framingham heart study dataset (Dawber et al., 1951) in Section 5.7.

### 3.5 Uncovering heterogeneity for causal discovery

For a response variable $Y$, consider the problem of finding the causal predictors $X^{PA_y}$, with $PA_Y \subseteq \{1, \ldots, d\}$, among a set of potential predictors $X^1, \ldots, X^d$. The method of invariant causal prediction (ICP) (Heinze-Deml et al., 2018; Peters et al., 2016; Pfister et al., 2018), for example, assumes that data are observed in different environments and that the causal mechanism for $Y$, given its causal predictors $PA_Y$ is invariant over the observed environments (see also Aldrich, 1989; Haavelmo, 1944; Pearl, 2009). This allows for the following procedure: For all subsets $S \subseteq \{1, \ldots, d\}$ one tests whether the conditional $Y|X^S$ is invariant. The hypothesis is true for the set of causal parents, so taking the intersection over all such invariant sets yields, with large probability, a subset of $PA_Y$ (Peters et al., 2016). Environments can, for example, correspond to different interventions on a node $X$. Using the concept of testing under distributional shifts, we can apply a similar reasoning even if no environments are available and one causal conditional is known instead.

Assume a causal model (e.g., a structural causal model, SCM, see Appendix A.2) over the variables $Y, X^1, \ldots, X^d$ and denote the causal predictors of $X'$ by $PA_x$. Assume further that there is a $j$ for which the conditional $q^*(x'|PA_x)$ is known. To infer the causal parents of $Y$ from a given data set, we split the data into two disjoint subsamples, indexed by $K \in \{1, 2\}$, and consider a shift which transforms the $K = 2$ sample into a new distribution, in which the conditional $q^*(x'|PA_x)$ has been changed to another conditional $p^*(x'|PA_x)$—this corresponds to a distribution generated by an intervention on $X$. We then take the two subsamples, the untransformed one, corresponding to $K = 1$, and the transformed one, corresponding to $K = 2$, as two ‘environments’ and apply the ICP methodology by testing whether the conditional $Y|X^S$ is invariant w.r.t. these two environments. That is, in the absence of ‘true heterogeneity’, we use the known conditional to artificially sample heterogeneity. Formally, for a candidate set $S \subseteq \{1, \ldots, d\}$ and an indicator variable $K$ indexing the two environments, we define the hypothesis

$$H_{0,S} := \{ P : Y \mid X^S_{|K=1} \equiv Y \mid X^S_{|K=2}, \quad p(x'|PA_x, k = 1) = q^*(x'|PA_x) \quad \text{and} \quad p(x'|PA_x, k = 2) = p^*(x'|PA_x) \}$$

and the shift factor $r(x', PA_x, k)$ similar to the one in Section 3.3. Here we slightly abuse notation, by using $K = 2$ to denote the transformed subsample. Naturally, the procedure extends
to $K > 2$. The distributional shift corresponds to an intervention on $X_i$ and it follows by modularity\(^6\) that $H_{0,PA_Y}$ is true. Therefore, the intersection over all sets for which $H_{0,S}$ holds trivially satisfies
\[ \bigcap_{S: H_{0,S} \text{ holds}} S \subseteq PA_Y, \]
where we define the intersection over an empty index set as the empty set. Our framework allows for testing such hypotheses from finitely many data (that were generated only using the conditional $q^*(x\mid PA_Y)$) and proves theoretical results that imply level statements for testing $H_{0,S}$. Such guarantees carry over to coverage statements for $\hat{S} := \cap_{S: H_{0,S} \text{ not rej.}} S$, that is, $\hat{S} \subseteq PA_Y$ with large probability.

### 3.6 Model selection under covariate shift

Consider the problem of comparing models in a supervised learning task when the covariate distribution changes compared to the distribution that generated the training data. Formally, let us consider an i.i.d. sample $D := \{(X_i, Y_i)\}_{i=1}^n$ from a distribution $q^*$, where $X_i \in \mathcal{X}$ are covariates with density $q^*(x)$ and $Y_i \in \mathcal{Y}$ is a label with conditional density $q^*(y|x)$. First, we randomly split the sample into two distinct sets, which we call training set $D_{\text{train}}$ and test set $D_{\text{test}}$. Let $\hat{f}_1: \mathcal{X} \rightarrow \mathcal{Y}$ and $\hat{f}_2: \mathcal{X} \rightarrow \mathcal{Y}$ be outputs of two supervised learning algorithms trained on $D_{\text{train}}$. In model selection under covariate shift (e.g., Quiñonero-Candela et al., 2009), we are interested in comparing the performance of the predictors $\hat{f}_1$ and $\hat{f}_2$ on a distribution $p^*$, where the covariate distribution is changed from $q^*$ to $p^*$, but the conditional $p^*(y|x) = q^*(y|x)$ remains the same. If we had an i.i.d. data set $D_{\text{test}}^{sh}$ from the shifted distribution $p^*$, we could compare the performances using a scoring function $S(D_{\text{test}}^{sh}, \hat{f})$ that for each of the predictors outputs a real-valued evaluation score. However, we only have access to $D_{\text{test}}$, which comes from $q^*$. Let us now assume that the shift from $q^*(x)$ to $p^*(x)$ is known. Existing methods use IPW to correct for the distributional shift (Sugiyama et al., 2007), which requires that the scoring function can be expressed in terms of an expectation of a single observation, such as the mean squared error. However, such a decomposition is not immediate for many scoring functions as for example the area under the curve (AUC). The framework of testing under distributional shifts allows for an arbitrary scoring function (as long as a corresponding test exists) while maintaining statistical guarantees. To this end, we define the hypothesis

\[ H_{0,\hat{f}_1,\hat{f}_2} := \{P: \mathbb{E}_{D_{\text{test}}} S(D_{\text{test}}^{sh}, \hat{f}_1) - S(D_{\text{test}}^{sh}, \hat{f}_2) \leq 0, \quad p(x) = p^*(x), p(y|x) = q^*(y|x)\}, \]

with the shift factor $r(x) := p^*(x)/q^*(x)$. Using data $D_{\text{test}}$ from $q^*$, the methodology developed below allows us to test this hypothesis $H_{0,\hat{f}_1,\hat{f}_2}$, that is, whether, in expectation, $\hat{f}_1$ outperforms $\hat{f}_2$ in the target distribution $p^*$, which includes the shifted covariate distribution. In practice, the densities $p^*(x)$ or $q^*(x)$ may not be given but one can still estimate these densities from data and apply our framework using (6).

### 4 Testing by resampling

In Section 3, we listed various problems that can be solved by testing a hypothesis about a shifted distribution. In this section, we outline several approaches to test a target hypothesis $r(Q^*) \in H_0$, see (1), using a sample $X_n$ from the observed distribution $Q^*$. We initially consider the shift $r$ known, and later show that asymptotic level guarantees also apply if $r$ is estimated sufficiently well from data.

Our approach relies on the existence of a hypothesis test $\varphi_m$ for the hypothesis $H_0$ in the target domain and applies this test to a resampled version of the observed data, which mimics a sample in the target domain. We show that—under suitable assumptions—properties of the original test $\varphi_m$ carry over to the overall testing procedure $\psi_m$ (of combined resampling and testing, as defined in (11)).

\[^6\] Formally, given an SCM, the statement follows from the global Markov condition (Lauritzen, 1996) in the augmented graph, including an intervention node with no parents that points into $X'$. 
This section is organised as follows. First, in Section 4.1, we propose a resampling scheme, which we show in Section 4.2 has asymptotic guarantees. In Section 4.3, we discuss how to sample from the scheme in practice and we describe a number of extensions in Section 4.4. In Section 4.5, we show that a simpler rejection sampling scheme can be used if stricter assumptions are satisfied.

### 4.1 Distinct replacement (DRPL) sampling

We consider the setting, where \( r(q)(x) \propto r(x)q(x) \) for a known shift factor \( r \); see (5). First, we draw a weighted resample of size \( m \) from \( X_n \) similar to the sampling importance resampling (SIR) scheme proposed by Rubin (1987) but using a sampling scheme DRPL (‘distinct replacement’) that is different from sampling with or without replacement. More precisely, we draw a resample \( (X_{i_1}, \ldots, X_{i_m}) \) from \( X_n \), where \( (i_1, \ldots, i_m) \in \{1, \ldots, n\}^m \) is a sequence of distinct\(^7\) values; the probability of drawing the sequence \( (i_1, \ldots, i_m) \) is

\[
\varpi(i_1, \ldots, i_m) \propto \prod_{\ell=1}^m r(X_{i_\ell}) \propto \prod_{\ell=1}^m \frac{r(q(X_{i_\ell}))}{q(X_{i_\ell})} \quad \text{if}(i_1, \ldots, i_m) \text{ is distinct and otherwise}.
\]

We provide an efficient sampling algorithm and discuss different sampling schemes in Section 4.3. We refer to \( (X_{i_1}, \ldots, X_{i_m}) \) as the target sample and denote it by \( \Psi_{\text{DRPL}}^m(X_n, U) \), where \( U \) is a random variable representing the randomness of the resample. If the randomness is clear from context, we omit \( U \) and write \( \Psi_{\text{DRPL}}^m(X_n) \). When \( m \) is fixed and \( n \) approaches infinity, the target sample \( \Psi_{\text{DRPL}}^m(X_n) \) converges in distribution to \( m \) i.i.d. draws from the target distribution \( r(Q^*) \); see Skare et al. (2003) for a proof for a slightly different sampling scheme. Based on our proposed resampling scheme, we construct a test \( \Psi_m \) for the target hypothesis (1) using only the observed data \( X_n \) by defining

\[
\Psi_m(X_n) := \varphi_m(\Psi_{\text{DRPL}}^m(X_n)),
\]

see also Algorithm 1.

**Algorithm 1**  
Testing a target hypothesis with known distributional shift and resampling

1. \( (i_1, \ldots, i_m) \leftarrow \text{sample from } \{1, \ldots, n\}^m \text{ with weights } (10) \) (see Appendix D)
2. \( \Psi_{\text{DRPL}}^m(X_n) \leftarrow (X_{i_1}, \ldots, X_{i_m}) \)
   
   return \( \Psi_m(X_n) := \varphi_m(\Psi_{\text{DRPL}}^m(X_n)) \)

We show in Section 4.2 that distinct sampling, \( \Psi_{\text{DRPL}} \), allows us to show guarantees without introducing regularity assumptions on the test \( \varphi_m \). The motivation for resampling without replacement comes from the fact that tests, as opposed to estimation of means, may be sensitive to duplicates; an extreme but instructive example is a test of the null hypothesis that no point mass is present in a distribution. A resampling test with large replacement size and possible duplicates would not be able to obtain level in such a hypothesis. Although we show in Appendix E that under stricter assumptions, sampling with replacement, that is using \( \Psi_{\text{REPL}} \), becomes asymptotically equivalent to using \( \Psi_{\text{DRPL}} \), this example highlights, that in the nonasymptotic regime, sampling duplicates may be harmful.

The resampling scheme \( \Psi_{\text{DRPL}} \) in (10) is similar, but not identical to what would commonly be called ‘resampling without replacement’ (\( \Psi_{\text{NO-REPL}} \)), where one draws a single observation \( X_{i_1} \), removes \( X_{i_1} \) from the list of candidates for further draws and normalises the remaining weights to reflect the absence of \( X_{i_1} \) (see also Section 4.3). \( \Psi_{\text{DRPL}} \) and \( \Psi_{\text{NO-REPL}} \) differ in the normalisation constants, and the normalisation constant in \( \Psi_{\text{DRPL}} \) is easier to analyse theoretically. This enables Lemma 1 in Appendix H, which describes the asymptotic behaviour of the mean and variance

\(^7\) We use ‘distinct’ and ‘nondistinct’ only to refer to the potential repetitions that occur due to the resampling \( (i_1, \ldots, i_m) \) and not due to potential repetitions in the values of the original sample \( X_n \).
of (10) as well as of the normalisation constant of (10). We consider $\Psi_{\text{DRPL}}$ a tool that enables simpler theoretical analysis of SIR methods; in practice it is plausible that using $\Psi_{\text{NG-REPL}}$ instead of $\Psi_{\text{DRPL}}$ will yield similar results, though we are not aware of any theory justifying this.

4.2 Pointwise asymptotic level and power

We now prove that the hypothesis test $\psi^r_n$ inherits the pointwise asymptotic properties of the test $\phi$ in the target domain. To do so, we require two assumptions: $m$ and $n$ have to approach infinity at a suitable rate, and we require the weights to be well-behaved. More precisely, we will make the following assumptions.

(A1). $m = m(n)$ satisfies $1 \leq m \leq n$, $m \to \infty$ and $m = o(\sqrt{n})$ for $n \to \infty$.

(A2). $E_Q [r(X)^2] < \infty$.

(A1) states that $m$ must approach infinity at a slower rate than $\sqrt{n}$. (A2) is a condition to ensure the weights are sufficiently well behaved, and is similar to conditions required for methods based on IPW, for example (Robins et al., 2000). If $r(x^A)$ only depends on a subset $A$ of variables, and $x^A$ takes finitely many values, (A2) is trivially satisfied for all $Q$. In the case of an off-policy hypothesis test such as the one described in Section 3.2, a sufficient but not necessary condition for (A2) to hold for $Q^*$ is that the policy $q^*(a|x)$ is randomised, such that there is a lower bound on the probability of each action. For a Gaussian setting, where $r$ represents a change of a conditional $q(x'|x')$ to a Gaussian marginal $p(x')$, we provide in Appendix I sufficient and necessary conditions under which (A2) is satisfied. If the hypothesis of interest is in the observed domain (see Section 2.4), we are usually free to choose any target density, so we can ensure that the tails decay sufficiently fast to satisfy (A2). In Section 5.1, we analyse the influence of (A1) and (A2) on our test holding level in the context of synthetic data. We now present the first main result which states that if $\alpha_\phi := \limsup_{k \to \infty} \mathbb{P}_P \left( \psi_k(Z_k) = 1 \right)$ is the asymptotic level of the test $\phi$ when applied to a sample $Z_k$ from $P^*$, then this is also the asymptotic level of the resampling test in Algorithm 1 when applied to a sample $X_n$ from $Q^*$. All proofs can be found in Appendix H.

Theorem 1 (Pointwise asymptotics—known weights). Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $r: \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r: \mathcal{X} \to [0, \infty)$ exists, satisfying $r(q)(x) = r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = r(Q)$. Let $\psi_k$ be a sequence of tests for $H_0$ and define $\alpha_\psi := \limsup_{k \to \infty} \mathbb{P}_P \left( \psi_k(Z_k) = 1 \right)$. Let $m = m(n)$ be a resampling size and let $\psi^r_n$ be the DRPL-based resampling test defined by $\psi^r_n(X_n) := \phi_m(\Psi_{\text{DRPL}}^r(X_n))$, see Algorithm 1. Then, if $m$ and $Q$ satisfy (A1) and (A2), respectively, it holds that

$$\lim_{n \to \infty} \mathbb{P}_Q(\psi^r_n(X_n) = 1) = \alpha_\psi.$$  

The same statement holds when replacing both $\limsup$’s with $\liminf$’s.

Theorem 1 shows that the rejection probabilities of $\psi^r$ and $\phi$ converge towards the same limit. In particular, the theorem states that if $\phi$ satisfies pointwise asymptotic level in the sense of (7), and (A2) holds for all $Q \in r^{-1}(H_0)$, then also $\psi^r$ satisfies pointwise asymptotic level (3). Similarly, because the statement holds for $P \notin H_0$, too, $\psi^r$ has the same asymptotic power properties as $\phi$.

We show in Theorem 1, that (A1) is sufficient to obtain asymptotic level of the rejection procedure. In fact, as we show in the following theorem, it is also necessary: If, $m$, $n$ approach infinity with $m \geq n^\ell$ for $\ell > \frac{1}{2}$, there exists a distribution $Q$, a shift $r$ and a sequence of tests $\psi_k$ such that (A2) is satisfied and $\alpha_\psi := \limsup_{k \to \infty} \mathbb{P}_P \left( \psi_k(Z_k) = 1 \right) < 1$ but the probability of rejecting the hypothesis on any resample of size $m$ converges to 1. This applies for any resampling scheme, including sampling with replacement.

Theorem 2 (In general, (A1) cannot be relaxed). Fix $\ell \in \{2, 3, \ldots\}$ and let $\Psi^m$ be any resampling scheme that outputs a (not necessarily distinct) sample of size
The grounds of Theorem 2, it is tempting to think that if one strengthens (A2) to
is relevant, for example, if the shift
Next, we consider the setting in which the shift factor is allowed to explicitly depend on
satisfies pointwise asymptotic level for the hypothesis
The following theorem states that even when estimating the weights, it is possible to obtain point
wise asymptotic level for the target hypothesis ( Pointwise asymptotics—estimated weights).

(A1'). \( m = m(n_2) \) satisfies \( 1 \leq m \leq n_2, m \to \infty \) and \( m = o(\min(n_1^{1/2}, n_2)) \) for \( n_1, n_2 \to \infty \),

(A2'). \( \mathbb{E}_Q[r_q(X_i)^2] < \infty \),

The following theorem states that even when estimating the weights, it is possible to obtain pointwise asymptotic level for the target hypothesis (1)—if the weight estimation works sufficiently well.

**Theorem 3** (Pointwise asymptotics—estimated weights). Consider a null hypothesis
\( H_0 \subseteq \mathcal{P} \) in the target domain. Let \( r : \mathcal{Q} \to \mathcal{P} \) be a distributional shift, satisfying \( r(q)(x) \propto r_q(x)q(x) \), see (6). Consider an arbitrary \( Q \in \mathcal{Q} \) and \( \mathcal{P} = r(Q) \).
Let \( \phi_k \) be a sequence of tests for \( H_0 \) and let \( \alpha_e := \limsup_{k \to \infty} \mathbb{P}_P(\phi_k(Z_k) = 1) \). Let \( \hat{r}_n \) be an estimator for \( r_q \) such that there exists \( a \in (0, 1) \) satisfying

\[
\limsup_{n \to \infty} \mathbb{E}_Q \left| \frac{\hat{r}_n(x)}{r_q(x)} - 1 \right| = 0,
\]

where the expectation is taken over the randomness of \( \hat{r}_n \). Let \( m = m(n_2) \) be a resampling size and let \( \psi_{n_1,n_2}^r := \phi_m(\Psi_{DRPL}^{r_{\hat{r}_n}(X,n)}) \) from Algorithm 3 in Appendix C. Then, if \( m \) and \( Q \)
satisfy (A1') and (A2'), respectively, it holds that

\[
\limsup_{n \to \infty} \mathbb{P}_Q(\psi_{n_1,n_2}^r(X_n) = 1) = \alpha_e.
\]

The same statement holds when replacing both lim sup's with lim inf's.

Theorem 3 shows that \( \psi_{n_1}^r \) converges to the same limit as \( \phi \). In particular, as for the case of known weights, if \( \phi \) satisfies pointwise asymptotic level and (A2') holds for all \( Q \in \tau^{-1}(H_0) \), then also \( \psi_{n_1}^r \) satisfies pointwise asymptotic level for the hypothesis \( \tau(Q^*) \in H_0 \) and \( \psi_{n_1}^r \) inherits asymptotic power properties, too.
4.3 Computationally efficient resampling with $\Psi_{\text{DRPL}}$

In Section 4.1, we propose a sampling scheme $\Psi_{\text{DRPL}}$, defined by (10), and in Section 4.2 we prove theoretical level guarantees when we resample the observed data using $\Psi_{\text{DRPL}}$. In this section, we display a number of ways to sample from $\Psi_{\text{DRPL}}$ in practice.

To do so, let $\Psi_{\text{REPL}}$ and $\Psi_{\text{NO-REPL}}$ denote weighted sampling with and without replacement, respectively, both of which are implemented in most standard statistical software packages. Though $\Psi_{\text{DRPL}}$ and $\Psi_{\text{NO-REPL}}$ both sample distinct sequences $(i_1, \ldots, i_m)$, they are not equal, i.e., they distribute the weights differently between the sequences (see Appendix D). We can sample from $\Psi_{\text{DRPL}}$ by sampling from $\Psi_{\text{REPL}}$ and rejecting the sample until the indices $(i_1, \ldots, i_m)$ are distinct, see Appendix D.1. In Proposition 2, we prove that under suitable assumptions, such as $m = o(\sqrt{n})$, the probability of drawing a distinct sequence already in a single draw approaches 1, when $n \to \infty$.

In some cases [though these typically only occur when (A1) or (A2) are violated, and our asymptotic guarantees do not apply], the above rejection sampling from $\Psi_{\text{REPL}}$ may take a long time to accept a sample. For these cases, we propose to use an (exact) rejection sampler based on $\Psi_{\text{NO-REPL}}$, which will typically be faster (since it has the same support as $\Psi_{\text{DRPL}}$). We provide all details in Section D.2.

If neither of the two exact sampling schemes for $\Psi_{\text{DRPL}}$ is computationally feasible, we provide an approximate sampling method that applies a Gibbs sampler to a sample from $\Psi_{\text{NO-REPL}}$; we refer to this scheme as $\Psi_{\text{DRPL-GIBBS}}$. Finally, one can simply approximate $\Psi_{\text{DRPL}}$ by a sample from $\Psi_{\text{NO-REPL}}$ – this is computationally faster, and leads to similar results in many cases. The details are provided in Section D.3. In practice, our implementation first attempts to sample from $\Psi_{\text{NO-REPL}}$ by (exact) rejection sampling, and if the number of rejections exceed some threshold, sampling without replacement is used instead.

Proposition 2 (mentioned above) has another implication. We prove that we can obtain the same level guarantee, when using $\Psi_{\text{REPL}}$ instead of $\Psi_{\text{DRPL}}$ (see Corollary 3 in Appendix D). This result, however, requires an assumption that is stronger than (A2). Intuitively, stronger assumptions are required for $\Psi_{\text{REPL}}$ because sampling with replacement is much more prone to experience large variance due to observations with huge weights.

4.4 Extensions

In this section, we discuss a number of extensions of the methodology and theory presented in the preceding sections.

4.4.1 Heuristic data driven choice of $m$

Resampling distinct sequences requires that we choose a resampling size $m$ that is smaller than the original sample size $n$. If $m$ is too large when sampling distinct sequences, it can happen that eventually there are no more points left that are likely under the target distribution. Consequently, the resampling procedure disproportionally often has to sample points that are very unlikely in the target distribution. This leads to the target sample being a poor approximation of the target distribution. Our theoretical results show that choosing a resampling size of order $o(\sqrt{n})$ avoids this problem, see Theorem 1.

However, this result is asymptotic and does not immediately translate into finite sample statements. Furthermore, in many cases also the requirement $m = o(\sqrt{n})$ is too strict, and asymptotic level can also be obtained by setting $m = o(n^a)$ for some $a \in (1/2, 1]$ (with the most extreme case being $P^* = Q^*$, where $a = 1$ can be applied). Since a larger $m$ typically results in increased power of the hypothesis test, we want to choose $m$ as large as possible while maintaining that the target sample still approximates the target distribution.

Consider the case where $r$ corresponds to changing $q^*(x'|x^C)$ to $p^*(x'|X^B)$. We can then test the validity of the resampling by testing whether the target sample matches the theoretical conditional. Specifically, for a fixed $m$, we can verify whether the conditional $X' \mid X^B$ in the resampled data $\Psi_{\text{DRPL}}(X_n)$ is close to the target conditional $p^*(x'|X^B)$ by a goodness-of-fit test $\chi^2(\Psi_{\text{DRPL}}(X_n)) \in (0, 1]$. If $m$ is chosen too large, the resampling is likely to include many points with small weights, corresponding to small likelihoods $p^*(x'|X^B)$, which will cause the goodness-of-fit test to reject the hypothesis that the target sample has the conditional $p^*(x'|X^B)$. 


We can use this to construct a data-driven approach to selecting $m$: For an increasing sequence of $m$'s, perform the goodness-of-fit test for several resamples $\kappa(\Psi_{\text{DRPL}}^{r,m}(X_n))_1, \ldots, \kappa(\Psi_{\text{DRPL}}^{r,m}(X_n))_K$. If $\frac{1}{K} \sum_{k=1}^{K} \kappa(\Psi_{\text{DRPL}}^{r,m}(X_n))_k$ is smaller than some predefined cut-off $q_t$, we accept $m$ as a valid target sample size. We then use the largest accepted $m$ as the resampling size in the actual test for the hypothesis of interest. We summarise the procedure for finding $m$ in Algorithm 4 in Appendix F and call this the GOF-heuristic for choosing $m$.

To avoid potential dependencies between the tuning of $m$ and the hypothesis test, we could use sample splitting. In practice, however, we use the entire sample, since the dependence between the tuning step and the final test in our empirical analysis appears to be sufficiently low such that the level properties of the final tests were preserved, see e.g., the experiment in Section 5.1.

If the target conditional $p^*(x' | X^B)$ is a linear Gaussian conditional density (i.e., $X' | X^B \sim N(\beta^T X^B, \sigma^2)$, for some parameters $\beta$, $\sigma$ under $p^*$) the goodness-of-fit test can be performed by using a linear regression and testing the hypothesis that the regression slope in the resample is $\beta$. For more complex conditional densities, one should prefer a test that has (pointwise asymptotic) power against a wide range of alternatives. Here, we propose to use the kernel conditional-goodness-of-fit test by Jitkrittum et al. (2020) to test that the resampled data $\Psi_{\text{DRPL}}^{r,m}(X_n)$ has the desired conditional.

4.4.2 Combining different resamples
The proposed procedure draws a random resample and hence a different conclusion of the test may be drawn if the procedure is repeated. To reduce this randomness, we may wish to repeat the resampling and testing several times, and combine the tests into a single test.

Repeating the procedure can also have a positive impact on sample efficiency: While Theorem 1 shows that $m = o(\sqrt{n})$ suffices for any distribution shift satisfying (A2), this choice of $m$ is in many cases too strict. An extreme case is when $P^* = Q^*$, where one could as well test the hypothesis in the $n$ observed points instead of the $m$ resampled points; here, using a single subsample of size less than $\sqrt{n}$ is not optimal in terms of power.

By repeated sampling, a larger part of the information contained in the data can be exploited. A difficulty in combining tests from several resamples is that the test statistics are dependent, since they are computed on random subsets of the same data set. Here, the strength of the dependence needs to be taken into account. Considering the following two corner cases may help for building intuition. If there are $m$ weights that are much larger than the remaining ones, the draws and the test statistics of different draws are mostly identical. In this case, repeated draws do not contain additional information about the null hypothesis. If all weights are equal, however, the test statistics are less dependent and thereby the different draws contain partially complementary information.

Several procedures exist for combining dependent tests or statistics (e.g., Liu & Xie, 2020; Rüschendorf, 1982; Vovk & Wang, 2020). Here, we propose to use the procedure proposed by Hartung (1999) that considers tests for a single hypothesis and only requires access to $p$-values $p_1, \ldots, p_k$ that are uniformly distributed under the null hypothesis. The procedure then transforms the $p$-values using the inverse Gaussian CDF $t_i = \Phi^{-1}(p_i)$, estimates the (pairwise) covariance of $t_1, \ldots, t_k$ and, taking these covariances into account, considers a weighted average of the probits $t_i$. Under the null hypothesis, this weighted average follows a standard normal distribution, which can be used to construct a combined $p$-value for the null hypothesis.

Conducting repeated tests can be combined with the heuristic in Section 4.4.1: We apply the heuristic to choose a resample size $m$ which holds level, and repeatedly test the hypothesis using this resample size, which we then combine into a single $p$-value. As we demonstrate empirically in Section 5.2, this often increases power compared to only testing the hypothesis once.

8 Concretely, since under the null hypothesis, $\kappa(\Psi_{\text{DRPL}}^{r,m}(X_n))_k$ is uniform, we chose $q_t$ to be the $\alpha$-quantile of the mean of $K$ uniform distributions for some $\alpha \in (0, 1)$, see Appendix F. Doing so, ensures that for a fixed $m$, under the null hypothesis of the resample having the intended conditional, the test has level $\alpha$.

9 This assumes that the pairwise covariance $\text{cov}(t_i, t_j)$ is constant for all $i \neq j$. This is satisfied in our case, since each $p_i$ is a result of the same test.
4.4.3 Finite-sample level guarantees

In addition to the asymptotic results presented in Section 4.2, we now prove that the hypothesis test \( \psi_m \) inherits finite-sample level if the test \( \varphi \) in the target domain satisfies finite-sample guarantees.

**Theorem 4** (Finite sample level—known weights). Consider a null hypothesis \( H_0 \subseteq \mathcal{P} \) in the target domain. Let \( \tau: Q \rightarrow P \) be a distributional shift for which a known map \( r: X \rightarrow [0, \infty) \) exists, satisfying \( \tau(q)(x) = r(x)q(x) \), see (5). Consider an arbitrary \( Q \in \mathcal{Q} \) and \( P = \tau(Q) \). Let \( m \) be a resampling size and let \( \varphi_m \) be a test for \( H_0 \) and define \( \alpha_v := P(\varphi_m(Z_m) = 1) \). Also let \( \psi_m \) be the DRPL-based resampling test defined by \( \psi_m(X_n) := \varphi_m(\Theta_{DRPL}(X_n)) \), see Algorithm 1. Then, if \( Q \) satisfies (A2), it holds that

\[
P_Q(\psi_m(X_n) = 1) \leq \inf_{\delta \in (0,1)} \left( \frac{\alpha_v}{1 - \delta} + \frac{V(n, m)}{V(n, m) + \delta^2} \right),
\]

where \( V(n, m) = \left( \begin{pmatrix} n \\ m \end{pmatrix} \right) \sum_{l=1}^{m} \left( \begin{pmatrix} m - l \\ m - l \end{pmatrix} \right) (E_Q[r(X^2)]^l - 1) \).

Thus, if \( E_Q[r(X^2)] \) is known, one can evaluate the finite-sample level of the DRPL-based resampling test for any choice of \( m \). We show in Appendix B that the term \( V(n, m) \) can be computed efficiently and such that numerical under- or overflows is avoided, even if \( n \) and \( m \) are so large that evaluating the individual terms \( \left( \begin{pmatrix} n \\ m \end{pmatrix} \right) \) and \( \left( \begin{pmatrix} n \\ l \end{pmatrix} \begin{pmatrix} m - l \\ m - l \end{pmatrix} \right) \) \( E_Q[r(X^2)]^l \) may cause under- or overflows. Given \( V(n, m) \), the minimisation problem on the right hand side can easily be implemented in numerical optimisers or solved explicitly for the minimal \( \delta \). Appendix B.1 contains plots of the upper bound for various values of the parameters.

4.4.4 Uniform level

The asymptotic level guarantees implied by Theorem 1 are pointwise, meaning we are not guaranteed the same convergence rate for all distributions \( Q \in \tau^{-1}(H_0) \). However, as the following theorem shows, if a uniform bound on the weights exists, i.e., \( \sup_{Q \in \tau^{-1}(H_0)} E_Q[r(X^2)] < \infty \), and the test \( \varphi \) has uniform asymptotic level, the overall procedure can be shown to hold to uniform asymptotic level.

**Theorem 5** (Uniform asymptotic level). Assume the same setup and assumptions as in Theorem 1. If additionally \( \sup_{Q \in \tau^{-1}(H_0)} E_Q[r(X^2)] < \infty \) and \( \limsup_{k \to \infty} \sup_{P \in H_0} P(P(\varphi_k(Z_k) = 1) \leq \alpha_v \), then

\[
\limsup_{n \to \infty} \sup_{Q \tau^{-1}(H_0)} P_Q(\psi_m(X_n) = 1) \leq \alpha_v,
\]

i.e., \( \psi_m \) satisfies uniform asymptotic level \( \alpha_v \) for the hypothesis \( \tau(Q^*) \in H_0 \).

---

\(^{10}\) Taking the derivative with respect to \( \delta \) and equating it to 0, the resulting equation can be rewritten to a polynomial equation of degree 4. One can then evaluate the right-hand side of (12) at the (at most 4) roots and additionally the boundary point \( \delta = 0 \), and use the one that yields the smallest bound.

If \( r \) is the identity mapping, i.e., \( Q^* = P \), then \( E_Q[r(X^2)] = 1 \), and for all \( m \), \( V(n, m) = 0 \). As one would expect, in that case Theorem 4 states that for any \( m \), \( P_Q(\psi_m(X_n) = 1) \leq \alpha_v \), that is, the probability of rejecting when applying \( \varphi \) to the resampled data is upper bounded by the probability of rejecting when applying \( \varphi \) directly to target data.

If \( P \neq Q^* \) then \( V(n, m) > 0 \), and for any \( m \) the right-hand side of (12) exceeds \( \alpha_v \). To control the level of the resampling test, say at a rate \( \alpha_v \), one can set the resample size \( m \) small enough such that the right-hand side of (12) is smaller than \( \alpha_v \).

We propose to use the largest \( m \) such that the right-hand side of (12) is bounded by \( \alpha_v \). In practice, we find that in many settings, the inequality (12) is not strict: the largest \( m \) such that the right-hand side (12) is bounded by \( \alpha_v \) is not close to being the largest \( m \) such that the left-hand side is bounded by \( \alpha_v \). Hence, for practical purposes, the scheme for choosing \( m \) proposed in Section 4.4.1 often returns larger values \( m \) while retaining level \( \alpha_v \) under the null hypothesis; we explore this further in Section 5.
4.4.5 Hypothesis testing in the observed domain
In Section 2.4, we argue that one can use our framework for testing under distributional shifts to test a hypothesis in the observed domain, too. Indeed, the results in Section 4.2 directly imply the following corollary (see Corollary 4 in Appendix H.8 for a more detailed version).

**Corollary 1** (Pointwise level in the observed domain). Consider hypotheses \( H_0^Q \subseteq Q \) and \( H_0^P \subseteq P \) and let \( \tau : Q \rightarrow P \) be a distributional shift such that \( \tau(H_0^Q) \subseteq H_0^P \). Under the same assumptions as in Theorem 1, if \( \phi \) is a test that satisfies pointwise asymptotic level in the target domain, \( \psi_n^\tau \) satisfies pointwise asymptotic level for the hypothesis \( Q^* \in H_0^Q \).

This corollary for example applies to the test in Section 3.1, where we test a hypothesis of (marginal) dependence in a shifted distribution \( \tau(P) \) although the hypothesis of interest is conditional independence in \( Q \). While the requirement in (A1) that \( m = o(\sqrt{n}) \) implies that we are testing the hypothesis in a smaller sample size, the hypothesis of marginal independence is also statistically simpler to test than a hypothesis of conditional independence.

This also means that our approach may come with computational benefits: We can resample the data into a smaller data set, where the signal of interest is more concentrated. For example, applying a kernel conditional independence test such as the one proposed by Zhang et al. (2011) can be computationally expensive when \( n \) is large. On the contrary, using a (marginal) kernel independence test such as the one by Gretton et al. (2008) is computationally simpler and for the proposed procedure only needs to be applied to a resample of size \( o(\sqrt{n}) \). In our experiments in Section 5.4, we employ both these tests and find that the procedure of resampling and testing for marginal independence is indeed orders of magnitude faster.

4.5 An alternative for uniformly bounded weights
In Section 4.1, we propose the ‘distinct replacement’ resampling scheme and show in Theorems 1 and 4 that this has finite and asymptotic level. The procedure requires (A2), that is that the weights have finite second moment.

We now consider the stricter assumption that the weights are globally bounded. Although this assumption is not met for most distributions that are not compactly supported, this is satisfied for example by distributions on finite state spaces. We show that, under this assumption, one can use a rejection sampler with finite sample guarantees.

Suppose that \( r(q)/x \propto r(x)q(x) \) and there exists a known \( M \in (0, \infty) \) such that \( \sup_x r(x) \leq M \). Given a sample \( X_n = (X_1, \ldots, X_n) \) of size \( n \) from \( Q^* \), we can use a rejection sampler that retains observations \( X_i \) with probability \( r(X_i)/M \) (and otherwise discards them) to obtain a sample from \( P^* = \tau(Q^*) \), and apply a hypothesis test \( \varphi_m \) to the rejection sampled data; see Algorithm 2.

**Algorithm 2** Testing a target hypothesis with known distributional shift and rejection sampling

| Input: | Data \( X_n \), hypothesis test \( \varphi_m \), shift factor \( r(x^\cdot) \) and bound \( M \). |
| 1: | for \( i = 1, 2, \ldots, n \) do |
| 2: | Sample \( U_i \) uniform on \( (0, 1) \) |
| 3: | if \( U_i > \frac{nX_i}{M} \) then |
| 4: | Discard \( X_i \) |
| return | \( \psi_n^\tau(X) = \varphi_m(X_{i_1}, \ldots, X_{i_n}) \) |

If \( \varphi_m \) has level guarantees when applied to data \( Z_k \) from \( P^* \), we can test the hypothesis \( \tau(Q^*) \in H_0 \) with the same level guarantee, since the rejection sampled data \( (X_{i_1}, \ldots, X_{i_m}) \) are i.i.d. distributed with distribution \( P^* \). We state this as a proposition.

**Proposition 1** (Finite level—bounded weights). Consider a null hypothesis \( H_0 \subseteq P \) in the target domain. Let \( \tau : Q \rightarrow P \) be a distributional shift for which a known map \( r : \mathcal{X} \rightarrow [0, \infty) \) exists, satisfying for all \( x \): \( \tau(q)(x) \propto r(x)q(x) \) and \( r(x) \leq M \). Consider an arbitrary \( Q \in Q \) and \( P = \tau(Q) \). Let \( \varphi_k \) be a sequence...
of tests for $H_0$ and assume there exist $\alpha_\phi \in (0, 1)$ such that for each $k \in \mathbb{N}$:

$$\alpha_\phi = \sup_k \mathbb{P}(\zeta_k = 1).$$

Let $\psi_n^\prime(X_n)$ be the rejection-sampling test defined in Algorithm 2. Then it holds that

$$\mathbb{P}_Q(\psi_n^\prime(X_n) = 1) = \alpha_\phi.$$

### 4.6 Statistical inference beyond testing

In this paper, we mainly focus on statistical testing in the target domain. However, by choosing $H_0$ as a singleton and properly defining $\phi_k$ (which is not required to be a test), the result of Theorem 1 is strong enough to imply that other types of statistical inference remain valid after resampling, too. We formulate this as a corollary.

**Corollary 2** (Confidence intervals, consistency, asymptotic distribution). Let $\tau: \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which a known map $r: \mathcal{X} \to [0, \infty)$ exists, satisfying $\tau(q)(x) \propto r(x)q(x)$, see (5). Consider an arbitrary $\mathcal{Q} \in \mathcal{Q}$ and $\mathcal{P} = \tau(Q)$. Let $g_k$ be an estimator or a confidence region in the target domain for a parameter $\theta(\mathcal{P})$. Let $m = m(n)$ be a resampling size and evaluate the estimator on the resampled data set, that is, $b_n^\prime(X_n) := g_m(\Psi_n^{\text{DRPL}}(X_n))$. Then, if $m$ and $Q$ satisfy (A1) and (A2), respectively, $b$ inherits properties like coverage, consistency or asymptotic normality from $g$. More precisely, we have the following three statements.

(i) Coverage of confidence regions. Let $\alpha \in [0, 1]$ be arbitrary. Then,

$$\liminf_{k \to \infty} \mathbb{P}_P(g_k(Z_k) \supseteq \theta(\mathcal{P})) \geq 1 - \alpha \Rightarrow \lim_{n \to \infty} \mathbb{P}_Q(b_n^\prime(X_n) \supseteq \theta(\mathcal{P})) \geq 1 - \alpha.$$

(ii) Consistency of estimator. Let $\varepsilon > 0$. Then, for any suitable norm $\|\cdot\|$,

$$\lim_{k \to \infty} \mathbb{P}_P(\|g_k(Z_k) - \theta(\mathcal{P})\| > \varepsilon) = 0 \Rightarrow \lim_{n \to \infty} \mathbb{P}_Q(\|b_n^\prime(X_n) - \theta(\mathcal{P})\| > \varepsilon) = 0.$$

(iii) Asymptotic distribution of estimator. Let $F$ be a cumulative distribution function, possibly depending on $\mathcal{P}$, and let $c \in \mathbb{R}$ be such that $F$ is continuous in $x$. Then,

$$\lim_{k \to \infty} \mathbb{P}_P(g_k(Z_k) \leq x) = F(x) \Rightarrow \lim_{n \to \infty} \mathbb{P}_Q(b_n^\prime(X_n) \leq x) = F(x).$$

Here, part (i) follows from Theorem 1 by choosing $\phi_k(Z_k) := 1_{(g_k(Z_k) \supseteq \theta(\mathcal{P}))}$, because the proof of Theorem 1 works with any function $\phi_k$ that takes values in $\{0, 1\}$. Part (ii) follows with $\phi_k(Z_k) := 1_{\|g_k(Z_k) - \theta(\mathcal{P})\| > \varepsilon}$ and part (iii) when choosing $\phi_k(Z_k) := 1_{(g_k(Z_k) \leq x)}$.

### 4.7 Choosing resampling parameters

We summarise the choices of parameters required to apply our framework in Table 1. Given a data set of size $n$, we need to specify a resampling strategy and a resampling size $m$. Additionally, in cases where the hypothesis of interest is in the observed distribution (see Corollary 4), we need to specify a target distribution.

If the weights are uniformly bounded, $\sup_k r(x) \leq M$ with known $M$, we propose to use the rejection sampler from Section 4.5, which does not require a choice of $m$ and has finite sample guarantees (see Proposition 1). If the weights are not uniformly bounded or $M$ is unknown, one can use the DRPL resampling scheme in Appendix D, which has asymptotic level guarantees. One can also use the NO-REPL resampling scheme (Appendix E), which is computationally faster, though sampling with replacement from a finite sample for many test statistics requires one to specify an effective sample size.
To choose the resampling size, we propose to use the *GOF-heuristic* described in Section 4.4.1, which increases \( m \) as long as a goodness-of-fit test is accepted. We propose to use conservative test by setting the rejection level high (10–30%), to ensure that the goodness-of-fit test has sufficient power to detect if the resampling did not work. One can also use the *asymptotic heuristic* \( m = \lfloor \sqrt{n} \rfloor \), which ensures asymptotic level but may not be appropriate for finite samples, or the \( m \) with finite sample level guarantees from Theorem 4 which, however, may be too conservative in practice. We recommend to repeat the resampling and testing procedure and use the combination test from Hartung (1999) whenever possible, as described in Section 4.4.2.

In applications, where the target distributions is not given, one also needs to choose the target distribution \( p^* \). For example in the application of conditional independence testing (Section 3.1), we are interested in converting a conditional distribution \( q^*(x|z) \) into a marginal distribution \( p(z) \). In such cases, we can choose the marginal to ensure that the weights are well behaved, yielding a better performance of our methods. Often, using the marginal distribution in the observed distribution \( p(z) = q^*(z) \) works well.

### 5 Experiments

We present a series of simulation experiments that support the theoretical results developed in Section 4.2 and analyse the underlying assumptions. We also apply the proposed methodology to the problems described in Section 3. A simulation experiment for model selection under covariate shift (see Section 3.6) can be found in Appendix G.2. Unless noted otherwise, the experiments use the \( \Psi_{DREPL} \) resampling scheme. Code that reproduces all the experiments is available at [https://github.com/nikolajthams/testing-under-shifts](https://github.com/nikolajthams/testing-under-shifts).

### 5.1 Exploring assumptions (A1) and (A2)

We explore the impact of violating either (A1), stating that \( m = o(\sqrt{n}) \), or (A2), stating that the weights must have finite second moment in the observational distribution. To do so, we apply the procedure discussed in Section 3.1 that reduces a conditional independence test \( X \perp Y | Z \) in

<table>
<thead>
<tr>
<th>Principle</th>
<th>Target distribution</th>
<th>Resampling scheme</th>
<th>Resampling size ( m )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Select ( p^* ) as close to ( q^* ) as possible in that it ensures that weights are well-behaved</td>
<td>We suggest to draw distinct observations to ensure valid inference.</td>
<td>Select ( m ) as large as possible s.t. resample remains sufficiently close to ( q^* ).</td>
<td></td>
</tr>
<tr>
<td>If target distribution ( p^* ) is not given by application, sometimes minimising variance of weights is possible; in CI testing, often replacing ( q^*(x</td>
<td>z) ) by ( q^*(x) ) works well</td>
<td></td>
<td>• Asymptotic theory requires ( m = o(\sqrt{n}) )</td>
</tr>
<tr>
<td>Optimise over feasible target distributions ( p^* ) to minimise ( \text{Var}_{Q^*}(r(X)) )</td>
<td>In general, use ( \text{DREPL} )</td>
<td>• Asymptotic heuristic: ( m = \lfloor \sqrt{n} \rfloor )</td>
<td></td>
</tr>
<tr>
<td></td>
<td>If computationally infeasible or ( n ) very large, use ( \text{NO-REPL} ) (assuming (A2’))</td>
<td>• GOF-heuristic: ( m ) as large as possible until GOF test rejects (Section 4.4.1) (make GOF as powerful as possible to maintain overall level)</td>
<td></td>
</tr>
</tbody>
</table>

• Improve power and reduce randomness by drawing multiple resamples and combining the tests (Section 4.4.2)

<table>
<thead>
<tr>
<th>Details</th>
<th>Summary of how to select the user-specified parameters of the resampling procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>† If target distribution ( p^* ) is not given by application, sometimes minimising variance of weights is possible; in CI testing, often replacing ( q^*(x</td>
<td>z) ) by ( q^*(x) ) works well</td>
</tr>
<tr>
<td>† If computationally infeasible or ( n ) very large, use ( \text{NO-REPL} ) (assuming (A2’))</td>
<td></td>
</tr>
<tr>
<td>† Asymptotic theory requires ( m = o(\sqrt{n}) )</td>
<td></td>
</tr>
<tr>
<td>† In general, use ( \text{DREPL} )</td>
<td></td>
</tr>
<tr>
<td>† GOF-heuristic: ( m ) as large as possible until GOF test rejects (Section 4.4.1) (make GOF as powerful as possible to maintain overall level)</td>
<td></td>
</tr>
<tr>
<td>† Improve power and reduce randomness by drawing multiple resamples and combining the tests (Section 4.4.2)</td>
<td></td>
</tr>
</tbody>
</table>

To choose the resampling size, we propose to use the *GOF-heuristic* described in Section 4.4.1, which increases \( m \) as long as a goodness-of-fit test is accepted. We propose to use conservative test by setting the rejection level high (10–30%), to ensure that the goodness-of-fit test has sufficient power to detect if the resampling did not work. One can also use the *asymptotic heuristic* \( m = \lfloor \sqrt{n} \rfloor \), which ensures asymptotic level but may not be appropriate for finite samples, or the \( m \) with finite sample level guarantees from Theorem 4 which, however, may be too conservative in practice. We recommend to repeat the resampling and testing procedure and use the combination test from Hartung (1999) whenever possible, as described in Section 4.4.2.

In applications, where the target distributions is not given, one also needs to choose the target distribution \( p^* \). For example in the application of conditional independence testing (Section 3.1), we are interested in converting a conditional distribution \( q^*(x|z) \) into a marginal distribution \( p(z) \). In such cases, we can choose the marginal to ensure that the weights are well behaved, yielding a better performance of our methods. Often, using the marginal distribution in the observed distribution \( p(z) = q^*(z) \) works well.
the goodness-of-fit test (horizontal dashed line) is rejected. (Right) Validation of (A2). Our procedure is run with regions show 95% quantiles of the rejection rates of the hypothesis test. The level seems to hold for rejection rates. (Left) Validation of (A1). We run the resampling with $\sigma = \sqrt{2}$ holds at 5%.

Figure 3. (All) Rejection rates from the experiments in Section 5.1. We replace the conditional distribution $q^*(z|x)$ with a marginal distribution $\phi_0(z)$. We perform a test for independence $X \perp Y$ in the target distribution, and plot the rejection rates. (Left) Validation of (A1). We run the resampling with $\sigma = 1$ and different sample sizes $m$. Shaded regions show 95% quantiles of the rejection rates of the hypothesis test. The level seems to hold for $m \leq \sqrt{n}$, the latter corresponding to the asymptotic rate (A1) (left of the dashed vertical line). Circles indicate the latter suggested by the middle plot and triangles the $m$ suggested by the finite-sample method described in Section 4.4.3—as expected, this is a conservative choice. The GOF-heuristic suggests an $m$ that indeed yields larger power. (Middle) $p$-values (dots) and average $p$-values (lines) when applying Algorithm 4 to choose $m$. We select $m$ (circled) from the first time, the goodness-of-fit test (horizontal dashed line) is rejected. (Right) Validation of (A2). Our procedure is run with different standard deviations $\sigma$ in the Gaussian target distribution $p_0(z) \phi_0(z)$. The dashed vertical line indicates the theoretical threshold of $\sqrt{6}$, see Section 5.1.

the observational domain to an unconditional independence test in the target domain. Specifically, we simulate $n = 10'000$ i.i.d. observations from the linear Gaussian model with

$$X := \varepsilon_X \quad Z := X + 2\varepsilon_Z \quad Y := \theta X + Z + \varepsilon_Y$$

for some $\theta \in \mathbb{R}$ and $\varepsilon_X, \varepsilon_Z, \varepsilon_Y \sim \mathcal{N}(0, 1)$ inducing a distribution $Q^*$ over $(X, Y, Z)$. We assume that the conditional distribution $q^*(z|x)$ is known and replace it with an independent Gaussian distribution $\phi_0(z)$ with mean zero and variance $\sigma^2$, breaking the dependence between $X$ and $Z$ in the target distribution.

We then perform a test for independence of $X$ and $Y$ in the target distribution using a Pearson correlation test. We do this both for $\theta = 0.4$ (where $X \perp Y \mid Z$ and ideally we reject the hypothesis) and for $\theta = 0$ (where $X \not\perp Y \mid Z$ and ideally we accept the hypothesis). Figure 3 shows the resulting rejection rates of the test, where we have repeated the procedure of simulating, resampling and testing (at level $\alpha = 0.05$) 500 (left) or 10'000 (right) times. In this experiment, the $\Psi_{\text{NO-REPL}}$ sampler is used, since the rejection samplers break as $m$ gets very large.

First, we test the impact of changing the resampling size $m$. For each simulated data set $X_n$ and each $m$, we resample 100 target data sets $\Psi^*_m(X_n)$, and compute the rates of rejecting the hypothesis. The shaded areas in Figure 3 (left) indicate 95% of the resulting trajectories, and the solid lines show one example simulation. Our theoretical results assume $m = o(\sqrt{n})$ and, indeed, the hypothesis rejects around 5% of simulations when $X \not\perp Y \mid Z$ for small $m$. As discussed in Section 4.4.1, (A1) may in some cases be too strict, and we observe that the 5% level is retained when $m$ moderately exceeds $\sqrt{n}$; but as $m$ grows larger, the level is eventually lost.

For the same example simulation $X_n$ as in the left plot, we also apply the GOF-heuristic for choosing $m$ as described in Section 4.4.1, and plot the resulting $p$-values in Figure 3 (middle). Since the data are Gaussian, we can perform the goodness-of-fit test by a simple linear regression analysis. For each $m$, we compute the average of the $p$-values (solid lines), and increase $m$ until the average $p$-value drops below the 5% quantile of the distribution of $\text{mean}(U_1, \ldots, U_l)$, where $U_1, \ldots, U_l$ are i.i.d. uniform random variables. The circles in the left and middle plot indicate the $m$ that is chosen by Algorithm 1 for this simulation. We observe in the left plot that the power of the test can be increased using the $m$ suggested by the middle plot, while the level approximately holds at 5%.
Second, we test the importance of \( (A2) \). For different \( \sigma \) (and fixed \( m = \sqrt{n} \)), we compute the weights \( \tau = \phi(Z_i)/q(Z|X) \), and in Figure 3 (right) we plot the rejection rates of the test statistic when \( X \rightarrow Y (\theta = 0.4) \) and \( X \not\rightarrow Y (\theta = 0) \). We show in Appendix I that \( (A2) \) is satisfied if and only if \( \sigma^2 < 2(\sigma_{\epsilon_{Z}}^2 - \sigma_X^2) \), where \( \sigma_X^2 \) is the variance of \( X \) and \( \sigma_{\epsilon_{Z}}^2 \) is the variance of the noise term in the structural assignment for \( Z \). In this experiment, it follows that \( (A2) \) holds if and only if \( \sigma < \sqrt{6} \). We observe that when \( \sigma \) exceeds the threshold of \( \sqrt{6} \) (vertical dashed line), the level eventually deviates from the 5% level. Furthermore, the power drops when \( \sigma \) approaches the threshold.

### 5.2 Combining repeated tests

In Section 4.4.2, we discussed combining multiple resamples, in order to both reduce the randomness of our resampling procedure, and to possibly get better finite sample performance, from using a larger part of the data.

In this experiment, we explore the effect of employing various combination tests, which allow for arbitrary dependencies between the tests. First, we consider the Cauchy combination tests (CCT) of Liu and Xie (2020). They assume that the test statistics are normally distributed, though they find that the outcome is not sensitive to this assumption in practice. We also consider the combination test of Hartung (1999), which transforms the p-values (assuming they are uniform under the null hypothesis) into a Gaussian distribution, and estimates the covariance between the tests in the transformed space. Lastly, we consider the combination procedure in Meinshausen et al. (2009), which combines p-values by taking quantiles of the empirical distribution (see also Rüger, 1978; Vovk & Wang, 2020).

We consider the task in Section 3.1 of testing conditional independence. We simulate data from both a linear and a nonlinear SCM, and test the hypothesis of conditional independence \( X \perp Y \mid Z \) both in a scenario where this null hypothesis is true and where it is false. In Figure 4, we plot the resulting rejection rates for various sample sizes \( n \), where for each \( n \), we sample \( m \) points, where \( m \) is selected either with the asymptotic heuristic \( m = \lfloor \sqrt{n} \rfloor \) or with the GOF-heuristic (see Section 4.4.1) which increases the resample size until the resample is rejected in a goodness-of-fit test at a 10% level.

In the linear SCM, where the weights are fitted using a (correctly specified) linear model, we observe that all three combination tests hold level and gain additional power over the single test. This is a well-behaved scenario, where the asymptotic heuristic is likely too conservative, and when used in conjunction with the asymptotic heuristic, the combination tests gain additional power by utilising more information from the sample. The gain in power of using combination tests is less pronounced for the GOF-heuristic, since more information is already used by the adaptive choice of \( m \).

In the nonlinear SCM, most types of combination tests also retain level, though in the case of the GOF-heuristic (which typically selects \( m \) more aggressively) the level is only attained for larger sample sizes or even not at all, for the test from Liu and Xie (2020). This could be because the goodness-of-fit test, used to decide how large the resample can be, may not have sufficient power in small sample sizes to detect when a resample does not follow the intended target distribution. This nonlinear SCM, where the weights are fitted by a generalised additive model (Hastie, 2017), is less well behaved than the linear SCM, and the weights are more degenerate; as a result, the tests are more correlated and the combination tests are not necessarily more powerful than the single test. This shows that while better power can be obtained by combining tests, it can also be lost in scenarios with degenerate weights, where the tests are strongly correlated.

### 5.3 Off-policy testing

We apply our method to perform statistical testing in an off-policy contextual bandit setting as discussed in Section 3.2. We generate a data set \( X_n = (Z_i, A_i, R_i) \) with dimensions \( d_Z = 3 \), \( d_A = d_R = 1 \), drawn according to the following data generating process:

\[
Z := \epsilon_Z \quad A \mid Z \sim q^*(A|Z) \quad R := \beta^\top A Z + \epsilon_R,
\]
where \( e_z \sim N(0, I) \) and \( e_R \sim N(0, 1) \), \( A \) takes values in the action space \( \{a_1, \ldots, a_L\} \), where \( L = 4 \), \( q^*(a|z) \) denotes an initial policy that was used to generate the data \( X_n \) and \( \beta_{a_1}, \ldots, \beta_{a_L} \) are parameters of the reward function corresponding to each action. A uniform random policy was used as the initial policy, i.e., for all \( a \in \{a_1, \ldots, a_L\} \) and \( z \in \mathbb{R}^3 \), \( q^*(a|z) = 1/L \).

The goal is to test hypotheses about the reward \( R \) if we were to deploy a target policy \( p^*(a|z) \) instead of the policy \( q^*(a|z) \). Here, we consider three hypotheses, namely one-sample test of means, two-sample test of difference in means and two-sample test of difference in distributions. We set the false positive rate to 5\% and use \( m = \sqrt{n} \) without the GOF-heuristic in all three experiments. Rejection rates are computed from 500 repeated simulations.

In the first experiment, we construct different target policies \( p^*_\delta(a|z) \). For \( \delta = 0 \), the target policy reduces to a uniform random policy and with increasing \( \delta \), the policy puts more mass on the optimal action (and thereby increasing the deviation from the initial policy). As \( \delta \to \infty \), the target policy converges to an optimal policy. More precisely, \( p^*_\delta(a|z) \) is a linear softmax policy, i.e., \( p^*_\delta(a|z) \propto \exp(\delta \beta_{a}^T z) \). We then apply our method to nonparametrically test whether \( \mathbb{E}_{P^\delta_{\delta}}(R) \leq 0 \) on the target distribution in which the policy \( p^*_\delta(a|z) \) is used. For \( \delta = 0 \), the expected reward is zero (here, the null hypothesis is true) and for increasing \( \delta \) the expected reward increases. To apply our methodology, we employ the Wilcoxon signed-rank test (Wilcoxon, 1992) in the target domain. Figure 5 (left) shows that for \( \delta = 0 \), our method indeed holds the correct level and eventually starts to correctly reject for increasing \( \delta \). For comparison, we include an estimate of the expected reward based on IPW.

In the second experiment, we use the same setup as in the first experiment, but now apply the two-sample testing method discussed in Section 3.3 to test whether \( R|_{K=1} \overset{\text{L}}{=} R|_{K=2} \), where \( K = 1 \) indicates a sample under the initial policy and \( K = 2 \) indicates a sample under a target policy. We consider two nonparametric tests, namely a kernel two-sample test based on the maximum mean discrepancy (MMD) (Gretton et al., 2012) (using the Gaussian kernel with the bandwidth chosen by the median heuristic Sriperumbudur et al., 2009) and the Mann–Whitney (M-W) U test (Mann & Whitney, 1947). Here, for \( \delta = 0 \), the two policies coincide and for \( \delta > 0 \), there is a difference in the expected reward. As shown in Figure 5 (middle), both tests are able to detect the difference. The M-W U test has more power than the MMD test.

In a third experiment, we construct different target policies \( p^*_\delta(a|z) \) by varying their effect on the variance of the reward distribution, while keeping the mean unchanged. More specifically, \( p^*_\delta(a|z) \) is a weighted random policy, i.e., \( p^*_\delta(a|z) \propto \delta a \) if \( a = a_1 \) and \( \propto 1 \) otherwise. This target policy yields the same expected reward as the initial policy (a uniform random policy), but yields a different variance of the reward. When \( \delta = 1 \), the target policy is the same as the initial policy, whereas the variance of the reward becomes smaller when \( \delta \) increases (in Figure 5 (right), \( \delta \) is rescaled.

---

**Figure 4.** Plotting rejection rates for the experiment for both a single test as well as three different combination tests, see Section 4.4.2.
We then apply the same two-sample testing methods used in the second experiment to test whether 

$$R_{|K=1} = R_{|K=2}.$$ 

This difference is not picked up by the M-W U test and this time, the MMD test has more power, see Figure 5 (right).

5.4 Testing a conditional independence with a complex conditional

In the setting of conditional independence testing, we now compare our method—when turning the problem into a test for unconditional independence as discussed in Section 3.1—to existing conditional independence tests. We sample \( n = 150 \) observations from the following structural causal model

\[
X := \text{GaussianMixture}(-2, 2) \quad Z := -X \cdot X + \varepsilon_Z \quad Y := \sin(Z) + \theta X^2 + \varepsilon_Y,
\]

inducing a distribution \( Q^* \), where \( \text{GaussianMixture}(-2, 2) \) is an even mixture (i.e., \( p = 0.5 \)) of two Gaussian distributions with means \( \mu_1 = -2, \mu_2 = 2 \) and unit variances, \( \varepsilon_Z, \varepsilon_Y \) are independent \( N(0, 4) \)-variables and \( \theta \in [0, 3/2], \tau \in \{1, 2\} \). Considering the conditional \( q^*(z|x) \) to be known, we apply our methodology for testing conditional independence \( X \perp Y \mid Z \) with a 5\% level and using the GOF-heuristic in Algorithm 4. To do so, we replace \( q^*(z|x) \) by a marginal density \( \phi(z) \), which is Gaussian with mean and variance set to the empirical versions under \( Q^* \). In the target distribution, we test for independence of \( X \) and \( Y \) using either a simple correlation test (CorTest) or a kernel independence test (HSIC) (Gretton et al., 2008). For comparison, we also conduct conditional independence tests in the observable distribution, using the CPT by Berrett et al. (2020), the generalised covariance measure (GCM) by Shah and Peters (2020) and a kernel conditional independence (KCI) by Zhang et al. (2011) (both using standard versions, without hyperparameter tuning). Our resampling methods use knowledge of the conditional \( q^*(z|x) \), which may be seen as an unfair advantage over the conditional independence tests. Therefore, we also apply our method with estimated weights, called HSICfit, where the conditional \( q^*(z|x) \) is estimated using a generalised additive model. Since CPT cannot exploit knowledge of \( q^*(z|x) \), we estimate the conditional \( q^*(x|z) \) to apply CPT. In this example, this is a more complex conditional than \( q^*(z|x) \) (Hoyer et al., 2008; Peters et al., 2014).

We repeat the experiment 500 times and plot the rejection rates in Figure 6 at various strengths of the edge \( X \rightarrow Y \). All instances of our method have the correct level, see rejection rates for \( \delta = 0 \). When \( \tau = 1 \), i.e., the direct effect \( X \rightarrow Y \) is linear, the power of our method approaches 100\% as the causal effect increases, albeit the conditional independence tests obtain power more quickly. When the direct effect is quadratic, CorTest and GCM have little or no power, as expected since

**Figure 5.** Off-policy statistical testing as described in Section 5.3. (Left) One-sample test for testing whether the mean under \( p^*_\delta(a|z) \) is less than or equal to 0, see Section 3.2. (Middle, Right) Two-sample tests for testing whether the reward under \( p^*_\delta(a|z) \) and \( p^\epsilon_\delta(a|z) \), respectively, has a different distribution than the reward under the initial policy, see Section 3.3. In all cases, the null hypothesis is true for \( \delta = 0 \). The target policies affect the mean of the reward in the left and middle plot, whereas they affect its variance in the right plot. The framework can be combined with nonparametric tests and thereby allows for detecting complex differences in the reward distribution when comparing two policies.
they are based on correlations (we believe that the slight deviation from 5% level in the left plot is due to very small sample sizes and the heuristic choice of \( m \)). KCI and HSIC have comparable power in the quadratic case, with our approach even obtaining slightly more power than KCI. We observe a slight level deviation for CPT (rejecting in 8.2% of cases with a confidence interval of (6.27%, 11.16%)), which may occur because the reverse conditional \( q^*(z|x) \) is harder to fit than \( q^*(x|z) \). Our approach has the additional benefit of low computational costs: conditional independence testing is usually a more complicated procedure than marginal independence testing and, furthermore, the marginal test is applied to a data set of size \( m \), which by (A1) is chosen much smaller than \( n \).

In Appendix G.3, we test the same conditional independences as Berrett et al. (2020) in the bike-share dataset by Fanaee-T and Gama (2014), and find that, when accepting hypotheses at 5% level, our conclusions coincide with those of Berrett et al. (2020) and Candès et al. (2018).

### 5.5 Conditional independence testing for right-censored data

We now apply our method for conditional independence testing to the setting of right-censored data. Consider a random vector \((X, Z, Y, C)\) with a joint density \( q^* \), where \( X \) is a vector of covariates, \( Z \) is a vector of control variables, \( Y \) is the time to an event of interest, and \( C \) is the time to the censoring event. In the setting, we do not directly observe \( Y \) and \( C \) but instead observe the right-censored time \( T = \min(Y, C) \) and the censoring indicator \( \delta = I\{T = Y\} \). The aim is to test the conditional independence \( X \perp Y \mid Z \) given an i.i.d. sample \( D \) of \((X, Z, T, \delta)\). To the best of our knowledge, there are no nonparametric conditional independence tests available in this setting (with nonempty conditioning set). Nonetheless, there exist nonparametric marginal independence tests (e.g., Fernández et al., 2021) that we can use to test \( X \perp Y \mid Z \) given the sample \( D \). We can, therefore, apply our method for conditional independence testing introduced in Section 3.1 to conduct a nonparametric test for \( X \perp Y \mid Z \) by testing \( X \perp Y \) in the target distribution in which the conditional \( q^*(x|z) \) is replaced by some marginal \( \phi(x) \).

To illustrate the use of our approach, we follow Fernández et al. (2021) and consider the colon dataset as our test bed. The data are from the study of adjuvant chemotherapy in patients with stage B/C colon cancer (Laurie et al., 1989; Moertel, 1995). Here, we consider the survival time (time to death) after the treatment as the event-time \( Y \), which was observed for around 49% of all the patients in the study, and consider testing whether the survival time \( Y \) is independent of the covariate \( X \) (the obstruction of colon by tumour) conditioned on the set of control variables \( Z \) (age, sex, and the extent of local spread).

We first analyse the dependency between the pairs \((X, Y)\), \((X, Z)\), and \((Z, Y)\). We use the Hilbert–Schmidt independence criterion (HSIC) (Gretton et al., 2008) to test for independence of the pair \((X, Z)\) and the kernel log-rank test for right-censored data (Fernández et al., 2021).
to test for independence of the pairs \((X, Y)\) and \((Z, Y)\). Table 2a reports the \(p\)-values of the (marginal) independence tests. At the confidence level of 95\%, the dependencies of all the pairs are significant. This motivates to investigate further whether the dependency between the obstruction of colon \(X\) and the survival time \(Y\) would still be significant when controlling for \(Z\), i.e., to test for \(X \perp Y \mid Z\).

To apply our method, we estimate the conditional density \(q^*(x|z)\) with logistic regression and compute the weights \(r = \hat{q}(x)/\hat{q}(x|z)\), where \(\hat{q}(x)\) is the empirical distribution of \(X\). The weights are then used to obtain the resample that mimics a sample from the target distribution in which the conditional \(q^*(x|z)\) is replaced by \(\hat{q}(x)\). The size of the resample is determined using our proposed GOF-heuristic with the threshold \(\alpha = 0.05\) (see Algorithm 4 in Appendix F). The heuristic selects \(m = 675\) from the total size of \(n = 929\), indicating that the two distributions are not far from each other. Table 2b reports the \(p\)-values of the independence tests in the resample. The result shows that the resampling successfully removes the dependency between \(X\) and \(Z\) in the resample (the \(p\)-value of the hypothesis \(X \perp Z\) is 0.386), while the dependency between \(X\) and \(Y\) remains significant. We can hence conclude that the dependency between the obstruction of colon \(X\) and the survival time \(Y\) is still significant when controlling for \(Z\) (age, sex and the extent of local spread).

### 5.6 Testing dormant independences

We now employ our method to test a dormant independence from observational data, as described in Section 3.4. We simulate data from a distribution \(Q^*\) that factorises according to the graph \(\mathcal{H}\) in Figure 2 and test the existence of the edge \(X^1 \rightarrow X^4\). As discussed by Shpitser and Pearl (2008), the presence of this edge cannot be tested by a conditional independence test, and instead we test marginal independence between \(X^1\) and \(X^4\) in the target distribution \(Q_{do}(X^3 := N)\), which can be obtained by applying our method using \(r_q(x^3, x^2) := q_{do}(X^3 := N)(x^3)/q(x^3|x^2)\), where \(q_{do}(X^3 := N)(x^3)\) is the density in the intervention distribution.

We conduct two experiments. In the first experiment, we consider binary random variables for the observables \(X^1, X^2, X^3\), and \(X^4\), while the hidden variable \(H\) is a discrete random variable with four possible values. We estimate \(q(x^3|x^2)\) by the empirical probabilities and use the empirical marginal distribution of \(X^3\) as a target distribution, i.e., \(p(x^3) = \hat{q}(x^3)\). In general, choosing the marginal distribution as a target distribution corresponds to a relatively small empirical variance of the weights \(r_q(x^3, x^2)\), see (A2), even though it may not always correspond to the minimum.\(^{11}\)

We employ Fisher's exact test to determine whether \(X^1\) and \(X^4\) in the interventional distribution are independent of each other. We compare our method to a more specialised method based on

\(^{11}\) In the special case of binary variables, when the density \(q(x^3, x^2)\) is known, one can compute the marginal which minimises the variance: it outputs one with probability

\[
q(x^3 = 1)/q(x^3 = 0) \cdot q(x^2 = 1)/q(x^2 = 0) + q(x^3 = 0)/q(x^3 = 1) \cdot q(x^2 = 0)/q(x^2 = 1).
\]

<table>
<thead>
<tr>
<th>Hypothesis</th>
<th>(P)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Original sample</td>
<td></td>
</tr>
<tr>
<td>(X \perp Y)</td>
<td>0.007</td>
</tr>
<tr>
<td>(X \perp Z)</td>
<td>0.028</td>
</tr>
<tr>
<td>(Z \perp Y)</td>
<td>0.006</td>
</tr>
<tr>
<td>(b) Resample</td>
<td></td>
</tr>
<tr>
<td>(X \perp Y)</td>
<td>0.002</td>
</tr>
<tr>
<td>(X \perp Z)</td>
<td>0.386</td>
</tr>
<tr>
<td>(Z \perp Y)</td>
<td>0.016</td>
</tr>
</tbody>
</table>

\(H\) in Figure 2 and test the existence of the edge \(X^1 \rightarrow X^4\). As discussed by Shpitser and Pearl (2008), the presence of this edge cannot be tested by a conditional independence test, and instead we test marginal independence between \(X^1\) and \(X^4\) in the target distribution \(Q_{do}(X^3 := N)\), which can be obtained by applying our method using \(r_q(x^3, x^2) := q_{do}(X^3 := N)(x^3)/q(x^3|x^2)\), where \(q_{do}(X^3 := N)(x^3)\) is the density in the intervention distribution.

We conduct two experiments. In the first experiment, we consider binary random variables for the observables \(X^1, X^2, X^3\), and \(X^4\), while the hidden variable \(H\) is a discrete random variable with four possible values. We estimate \(q(x^3|x^2)\) by the empirical probabilities and use the empirical marginal distribution of \(X^3\) as a target distribution, i.e., \(p(x^3) = \hat{q}(x^3)\). In general, choosing the marginal distribution as a target distribution corresponds to a relatively small empirical variance of the weights \(r_q(x^3, x^2)\), see (A2), even though it may not always correspond to the minimum.\(^{11}\)

We employ Fisher’s exact test to determine whether \(X^1\) and \(X^4\) in the interventional distribution are independent of each other. We compare our method to a more specialised method based on

\(^{11}\) In the special case of binary variables, when the density \(q(x^3, x^2)\) is known, one can compute the marginal which minimises the variance: it outputs one with probability

\[
q(x^3 = 1)/q(x^3 = 0) \cdot q(x^2 = 1)/q(x^2 = 0) + q(x^3 = 0)/q(x^3 = 1) \cdot q(x^2 = 0)/q(x^2 = 1).
\]
In the second experiment, we consider a linear Gaussian SCM. We estimate the conditional $q(x^3|x^2)$ by a linear regression and, as before, use the empirical marginal distribution of $X^3$ as a target distribution. We then test for independence between $X^1$ and $X^4$ in the interventional distribution using a simple correlation test. Since the distribution is jointly Gaussian (with linear functions) and satisfies the ‘bow-free’ condition (Brito & Pearl, 2002), there is a specialised, nontrivial likelihood procedure for model selection that we can compare with: We perform maximum-likelihood estimation as suggested by Drton et al. (2009) and use the penalty from Nowzohour et al. (2017) to score the graphs $G$ and $H$.

In both experiments, we consider two strategies for choosing the resampling size $m$: (1) $m = \sqrt{n}$ and (2) the GOF-heuristic (see Algorithm 4). The resulting rejection rates over 500 repeated experiments, for several sample sizes, are shown in Figure 7. Our method identifies both the absence and presence of the causal edge $X^1 \rightarrow X^4$ in both the binary and the Gaussian setting. Our method has the desired level (except for the Gaussian case when $m = 100$) and increasing power as sample size increases. Yet, as one would expect, the score-based approaches, which are tailored to each model class, have more power to detect the absence of the edge (though in the binary case, the level of the test does not seem to hold exactly when the sample sizes are small). Compared to $m = \sqrt{n}$, the choice of $m$ with the GOF-heuristic yields larger test power without sacrificing too much the level of the test (although the level is violated for small sample sizes in the Gaussian setting). Our method can readily be applied to nonlinear and non-Gaussian settings, too, see Appendix G.4.

### 5.7 Testing front-door assumptions

We apply our method to testing conditional independence under interventions to test the front-door assumptions as proposed by Bhattacharya and Nabi (2022) on the Framingham heart study dataset (Dawber et al., 1951). We revisit this dataset and apply our proposed procedure for choosing the resample size $m$ (see Section 4.4.1) and the method for combining different resamples (see Section 4.4.2). The Framingham heart study is a longitudinal epidemiology study of the risk factors for cardiovascular disease. Here, we consider a similar setup as Bhattacharya and Nabi (2022) and consider testing the front-door assumptions for estimating the effect of smoking (treatment $A$) on the development of coronary heart disease (outcome $Y$). The hypertension condition is considered as a mediator $M$ and the past history of hypertension as an anchor $Z$. In addition, the set of covariates $C$ containing age, sex, and BMI are included as control variables. Figure 8 (top left) illustrates the assumed underlying causal graph. To apply the front-door adjustment for estimating the causal effect of $A$ on $Y$, one major assumption is that there is no direct edge from $A$ to $Y$. Under...
faithfulness (Pearl, 2009), this assumption can be verified by testing $Z \perp Y \mid C$ in the interventional distribution in which $M$ is replaced by some marginal, see Figure 8 (bottom right). We apply our resampling approach by first estimating the conditional $\hat{q}(m|a, z, c) \approx q^*(m|a, z, c)$ with logistic regression and computing the weights $r = \hat{q}(m)/\hat{q}(m|a, z, c)$, where $\hat{q}(m)$ is the empirical distribution of $M$. The resample size $m$ is chosen using the GOF-heuristic with $\alpha_e = 0.05$ (see Algorithm 4).

In addition, we compute $p$-values from multiple resamples and combine them as discussed in Section 4.4.2.

Figure 8 (right) presents $p$-values of the (conditional) independence tests in the original sample (a) and in the resample (b). We use the kernel conditional independence (KCI) test (Zhang et al., 2011) and the Hilbert–Schmidt independence criterion (HSIC) test (Gretton et al., 2008) for the hypotheses $Z \perp Y \mid C$ and $M \perp (A, C, Z)$, respectively. The result suggests that the front-door assumption is plausible (the $p$-value of the hypothesis $Z \perp Y \mid C$ is 0.263). (Instead of a conditional independence test under intervention, Bhattacharya and Nabi (2022) use a slightly more involved testing procedure and come to the same conclusion.) Moreover, our approach successfully eliminates the effect from $(Z, A, C)$ on $M$ in the resample (the $p$-value of the hypothesis $M \perp (A, C, Z)$ is 0.295).

### 5.8 Comparison to IPW

Inverse probability weighting (IPW) allows us to test simple hypotheses such as $\mathbb{E}_P[f(X)] = c$ for some constant $c \in \mathbb{R}$ and a given function $f$. If data $Z_m$ sampled from the target distribution $P^*$ are available, we could test the hypothesis using the test statistic $\frac{1}{m} \sum_{i=1}^m f(Z_i)$. If, instead, data are available from a different observable distribution $Q^*$, we can estimate the corresponding test statistic in the target domain using the test statistic

$$T(X_n) := \frac{1}{n} \sum_{i=1}^n \tilde{r}(X_i)f(X_i),$$

where $\tilde{r}(X_i)$ is the normalised versions of the shift factor $r$ (elsewhere we do not require $r$ to be normalised). If $r(X_i)f(X_i)$ has finite second moment in $Q$, then $T(X_n)$ is asymptotically normal with mean $\mathbb{E}_P[f(X_1)]$ and variance $\sigma^2 := \sqrt{Q(r(X_1)f(X_1))}/\sqrt{n}$, and one can construct a (1 − $\alpha$) confidence interval as

$$[T(X_n) - z_{\alpha/2} \cdot \sigma/\sqrt{n}, T(X_n) + z_{\alpha/2} \cdot \sigma/\sqrt{n}],$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile from the standard normal distribution.

To compare our approach to the IPW approach, we simulate data ($n = 100$) from the following structural equation model:

$$X^1 := 1 + \varepsilon_{X^1}, \quad X^2 := X^1 + \varepsilon_{X^2}, \quad X^3 := X^2 - X^1 + \varepsilon_{X^3},$$
with \(\varepsilon_{X^1} \sim \mathcal{N}(0, 3)\), \(\varepsilon_{X^2} \sim \mathcal{N}(0, 4)\) and \(\varepsilon_{X^3} \sim \mathcal{N}(0, 1)\). In this model, the mean of \(X^3\) is \(\mathbb{E}_Q[X^3] = 0\).

We consider the distributional shift corresponding to the intervention \(\text{do}(X^2 := \mu + \bar{\varepsilon}_{X^2})\) with \(\bar{\varepsilon}_{X^2} \sim \mathcal{N}(0, 1)\), where \(X^3\) has mean \(\mathbb{E}_P[X^3] = \mu - 1\), and test the hypothesis \(\mathbb{E}_P[X^3] = 0\) for various \(\mu\) using both our resampling approach (with \(m\) chosen according to the GOF-heuristic in Algorithm 4) and the IPW-based confidence intervals. Since IPW is sensitive to degenerate weights, we also use a ‘clipped IPW’, where we truncate the 10 largest weights at the 10th largest value (see e.g., Cole & Hernán, 2008).

Ideally, we accept the hypothesis for \(\mu = 1\) and reject the hypothesis for all other \(\mu\). The larger \(\mu\) becomes, the easier it should be to reject the hypothesis \(\mu = 1\), if target data are available. At the same time, since the target distribution is a Gaussian distribution centred at \(\mu - 1\), as \(\mu\) increases, the weights get increasingly degenerate, because the weights of the data points with the largest numerical values \(X^2\) dominate the weights of all other data points.

We observe in Figure 9 that all methods have the correct level at 5% (when \(\mu = 1\)) and approximately the same power for small \(\mu\). As \(\mu\) grows, the plain IPW loses its power, due to weight degeneracy. Both the clipped IPW and our resampling approach do not suffer from this issue, with power approaching 1, even as weights get increasingly degenerate. This experiment indicates that our method may share some of the robustness to degenerate weights that is known from clipped IPW, and at the same time is able to estimate more complex test statistics, that cannot be estimated using IPW.

### 5.9 Resampling for heterogeneity to identify causal predictors

In Section 3.5, we propose to use our resampling approach to create heterogeneous data and apply ICP (Peters et al., 2016) to estimate (a subset of) the causal predictors \(X^{PA_2}\) of a response variable \(Y\), even when no environments are given. We now illustrate that this approach can infer casual relationships that would not be detectable using conditional independence statements. We therefore generate \(n = 1'000\) i.i.d. observations of \(Y, X^1, X^2, X^3, X^4\) according to a linear Gaussian SCM with the graphical representation given in Figure 10 (left). Furthermore, we assume that the conditional distribution \(q^*(x^2|x^1, x^3)\) is known (instead, one could also assume that \(PA_2 = \{1, 3\}\) is known and estimate the conditional). As described in Section 3.5, we now generate two environments by splitting the observational distribution into two, and modify one of them based on a distributional shift. Specifically, we split the data into two disjoint subsets \(K = 1\) and \(K = 2\) and then resample \(m = 30\) (approximately \(\sqrt{n}\)) observations from \(K = 2\) under the distributional shift generated by replacing the conditional \(q^*(x^2|x^1, x^3)\) with the target distribution \(p^*(x^2|x^1, x^3)\) (which flips the sign of the dependence on \(x^3\)). The precise data generating process is described in Section G.5. This results in a data set with \(n + m\) observations from two environments. We then apply ICP to the joint data from both environments and output the following estimate of the causal predictors:

\[
\hat{S} := \bigcap_{S: H_0,S \text{ accepted}} S,
\]
Here, $H_{0,5}$ is the hypothesis defined in Section 3.5. We use the InvariantCausalPrediction R-package for this experiment, which tests $H_{0, S}$ using a Chow test (Chow, 1960). We repeat the experiment 500 times and report in Figure 10 (right) how many times each set $S$ is output. As an oracle benchmark, we also report the corresponding frequencies when we sample the target distribution directly, instead of resampling it (in particular we use the same total sample size $m + n$). Our method frequently returns the invariant set $\{2, 3\}$ and holds the predicted coverage guarantee: in only 0.2% of the cases, the estimated set is not a subset of $\{2, 3\}$. The output of the method is guaranteed to be (with large probability) a subset of the set of true causal predictors, but depending on the type of heterogeneity, the method may output the empty set. e.g., if the true (unknown) underlying graph equals $X^4 \rightarrow Y \rightarrow X^1 \rightarrow X^2 \leftarrow X^3$, then (for the same experiment), both ICP based on the resampled data and ICP based on the true target distribution always output the empty set.

The difference between the oracle method and the resampling method (see Figure 10) indicates that the resampled distribution does not equal the target distribution. Indeed, in some regions where the target density has substantial mass, there are no data points that can be sampled. This, however, does not show any effect on the level of the overall procedure. Thus, in the resampled data the conditional distribution of $X^2$, given $X^1$ and $X^3$ differs from $q^*(x^2|x^1, x^3)$ (even though it does not equal the target conditional $q^*(x^2|x^1, x^3)$). We hypothesise that the result is therefore similar to choosing a different target distribution in the first place. Indeed, when changing $p^*(x^1, x^2, x^1)$ to match the data support, the set frequencies of the oracle version closely match the resampled version.

### 5.10 Comparing to double machine learning methods in treatment effect estimation

We consider a treatment effect estimation setup, where for a binary treatment $D$, observed founders $X$ and continuous outcome $Y$ we have that

$$P_{Q^*}(D = 1) = m_0(X) \quad Y = g_0(D, X) + \varepsilon_Y \quad \mathbb{E}_{Q^*}[\varepsilon_Y|X, D] = 0$$

for some unknown functions $m_0$ and $g_0$. We consider a setting, where $X$ is 20-dimensional, $m_0$ is a sigmoid function and $g_0$ is either a linear function or a complex, nonlinear function. We then consider the hypothesis (in the observed distribution) that the average treatment effect (ATE) is zero, where the ATE is given by $\theta_0 = \mathbb{E}_{Q^*}[g_0(1, X)] - \mathbb{E}_{Q^*}[g_0(0, X)]$. Semi-parametrically efficient estimation of $\theta_0$, using doubly robust methods that model both $m_0$ and $g_0$ is possible if estimators of these converge sufficiently fast (Chernozhukov et al., 2018; Robins & Rotnitzky, 1995; Robins et al., 1994).

In this experiment, we compare our resampling methodology to a doubly robust method, where for the latter we fit a logistic regression model to estimate $m_0$ and either a random forest (nonlinear case) or a linear regression (linear case) to estimate $g_0$, using the DoubleML package in Python.
We then test the hypothesis \( \theta_0 = 0 \) by checking whether a 95% confidence interval contains zero.

To apply our methodology to this problem, we also fit a logistic regression model to estimate the conditional \( q^*(d|x) \) and replace this conditional with a marginal target probability \( p(d) \), which matches the empirical marginal probability of treatment, \( q^*(d) \). We then apply a t-test to the resample (Student, 1908), to test the hypothesis of \( \theta_0 = 0 \) at a 5% level using the combination test from Hartung (1999), as discussed in Section 4.4.2. We apply this method both with the asymptotic heuristic (\( m = \lfloor \sqrt{n} \rfloor \)) and with the GOF-heuristic, where we chose \( m \) as large as possible while still accepting a goodness-of-fit test that the resample follows the target distribution (Section 4.4.1) at a 20% level.

We repeat the experiment 300 times and in Figure 11 we plot the rejection rates for the hypothesis both in a scenario where \( \theta_0 = 0 \) (no effect present) and where the ATE is nonzero (effect present). Our procedure, using either heuristic, and the double ML procedure satisfy the desired 5% level for all sample sizes. When the true outcome model \( g_0 \) is a complex function, our procedure with the GOF-heuristic outperforms the power of the double ML method; our resampling procedure does not model \( g_0 \), and so is unaffected by how difficult it is to estimate \( g_0 \). On the contrary, when \( g_0 \) and \( m_0 \) are both simple functions, the double ML approach (which we correctly specify by using a linear and a logistic regression) attains more power than our procedure. The GOF-heuristic has more power than the asymptotic heuristic in both the linear and the nonlinear setting, as one would expect since the GOF-heuristic use larger resampling sizes \( m \) to test the hypothesis.

**6 Conclusion and future work**

We formally introduce statistical testing under distributional shifts and illustrate that it can be applied in a diverse set of areas such as contextual bandits, conditional independence testing, and causal inference. We provide a general testing procedure based on weighted resampling and prove pointwise asymptotic level guarantees under mild assumptions. Our simulation experiments underline the usefulness of our method: It is able to test complicated hypotheses, such as dormant independences—for which to-date no test with provable level guarantees exists—and can be applied to test complex hypotheses in off-policy testing or covariate shift. The framework is competitive even in some of the problems, where more specialised solutions exist. Its key strength is that it is very easy to apply and can be combined with any existing test making it an attractive go-to method for complicated testing problems.

We believe that several directions would be worthwhile to investigate further. In many of the empirical experiments, the requirement that \( m = o(\sqrt{n}) \) seems too strict and can be relaxed, see
also Section 4.4.1. As discussed in Section 4.2, we hypothesise that under further restrictions on the weights or the test statistics, the assumption for the theoretical results can be relaxed. Bickel et al. (2012) consider the \( \binom{n}{m} \) bootstrap, which resamples distinct sequences without weights, and show that under mild assumptions, bootstrap estimates converge if \( m = o(n) \). Further work is required to extend this to the case of weighted samples.

We show in Section 4.4.4 that the main convergence result, Theorem 1, can be extended to uniform level, if we make uniform assumptions on the target test, \( \varphi \), and that the weights are uniformly bounded over \( \tau^{-1}(H_0) \). In many model classes, the latter assumption may be too strict, and a better understanding of necessary conditions would help.

While Theorem 3 provides guarantees when \( r(x^A) \) is unknown, the theorem requires guarantees on the relative error \( \hat{r}(x^A)/r(x^A) \). A more natural guarantee would be on the absolute error \( |\hat{r}(x^A) - r(x^A)| \), and we hope further work can shed light on the appropriate conditions (such as model classes \( p \) and \( q \) or properties of the estimator \( \hat{r} \)) to achieve such a guarantee.

Resampling distinct sequences is less prone to weight degeneracy than IPW or resampling with replacement, but in setups with well-behaved weights this may come at a cost of power when resampling only \( m \ll n \) points. Resampling nondistinct sequences share many similarities with IPW (for fixed \( n \), expectations of \( \Psi_{\text{REPL}} \) converge to the IPW estimate when \( m \to \infty \)), but additionally benefits from the ability to test hypotheses where the test statistic cannot be written as an average over the data points (see Section 5.8). Further investigation of the differences between the sampling schemes and benefits and disadvantages in comparison to IPW is needed.

Our methodology considers the setting where we only observe data \( X_n \) from the distribution \( Q^* \). If additionally a sample \( Z_n' \) from the target distribution \( P^* \) is already available, one can combine the two data sets, to get a larger approximate sample from \( P^* \), a problem known as ‘domain adaptation’ in the literature (Finn et al., 2017). In particular, if \( Z_n' \) is also available, one could perform the testing on the combined data set \((\Psi(X_n), Z_n')\). We believe that similar theoretical guarantees can be proved.

When testing for a hypothesis in the observed domain, we often have the freedom to choose a target distribution which could help us improve the performance of our test (as discussed in Section 2.4). In the experiments, e.g., Sections 5.3 and Section 5.4, we choose the target distribution that matches certain marginals, which often helps to minimise the variance of the weights. Another possibility is to choose a target distribution such that the alternative becomes easier to detect which can be achieved by minimising the \( p \)-value of the test with respect to the choice of the target distribution.

**Acknowledgments**

We thank Peter Rasmussen for valuable ideas about the combinatorics, Mathias Drton for helpful discussions on hidden variable models, Tom Berrett for insightful comments during a discussion of an earlier version of this paper, Rohit Bhattacharya for providing us the code for preprocessing the Framingham heart study data and the members of NT’s PhD committee, Susanne Ditlevsen, Thomas Richardson, and Rajen Shah for helpful comments.

Conflict of interest: None declared.

**Funding statement**

N.T., S.S., and J.P. were supported by a research grant (18968) from VILLUM FONDEN and J.P. was, in addition, supported by the Carlsberg Foundation. N.P. was supported by a research grant (0069071) from Novo Nordisk Fonden.

**Data availability**

The data that support the finding of this study are openly available at https://github.com/nikolajthams/testing-under-shifts.
Appendix A: Further comments on the framework

A.1 Forward and backward shifts, \(\tau\) and \(\eta\)

In this paper, as mentioned in Section 2.1, we take the starting point that \(Q^*\) is observed, and view \(P^* = \tau(Q^*)\) as a shifted version of \(Q^*\). One could instead suppose that we started with a distribution of interest \(P^*\), from which no sample is available, and then construct a map \(\eta\) such that \(Q^* = \eta(P^*)\) is a distribution which can be sampled from in practice. If \(\tau\) and \(\eta\) are invertible, the two views are mathematically equivalent, but if not, there is a subtle difference; the corresponding level guarantees take a supremum either over \(Q \in \{ Q' | \eta^{-1}(Q') \cap H_0 \neq \emptyset \}\) (\(\eta\) view) or over \(Q \in \{ Q | \tau(Q') \in H_0 \}\) (\(\tau\) view). To see this, we first start with the (natural) level guarantee from the \(\tau\) view: \(\sup_{P \in H_0} \mathbb{P}^\tau (\psi_n(X_n, U) = 1) \leq \alpha\). We then have

\[
\sup_{P \in H_0} \mathbb{P}_{\eta(P)} (\psi_n(X_n, U) = 1) \leq \alpha
\]

\[
\Leftrightarrow \sup_{Q \in \eta(H_0)} \mathbb{P}_Q (\psi_n(X_n, U) = 1) \leq \alpha
\]

\[
\Leftrightarrow \sup_{Q \in \{ Q | \tau(Q') \in H_0 \}} \mathbb{P}_Q (\psi_n(X_n, U) = 1) \leq \alpha.
\]

If, alternatively, we start with the level guarantee from the \(\eta\) view, we find

\[
\sup_{P \in H_0} \mathbb{P}_{\eta(P)} (\psi_n(X_n, U) = 1) \leq \alpha
\]

\[
\Leftrightarrow \sup_{Q \in \eta^{-1}(H_0)} \mathbb{P}_Q (\psi_n(X_n, U) = 1) \leq \alpha
\]

\[
\Leftrightarrow \sup_{Q \in \{ Q' | \tau(Q') \in H_0 \}} \mathbb{P}_Q (\psi_n(X_n, U) = 1) \leq \alpha.
\]

Comparing the last two lines yields the claim.

A.2 Example: interventions in causal models

One example of a distributional shift \(\tau\) is the case where \(\tau\) represents an intervention in a structural causal model (SCM) over \(X^1, \ldots, X^d\) (Pearl, 2009). An SCM \(S\) over \(X^1, \ldots, X^d\) is a collection of structural assignments \(f^1, \ldots, f^d\) and noise distributions \(Q_{N^1}, \ldots, Q_{N^d}\) such that for each \(j = 1, \ldots, d\), we have \(X^j := f^j(PA^j, N^j)\). Here, the noise variables \(N^j\) are assumed to be jointly independent. The sets \(PA^j \subseteq \{ X^1, \ldots, X^d \} \setminus \{ X^j \}\) denote the causal parents of \(X^j\). The induced graph over \(X^1, \ldots, X^d\) is the graph obtained by drawing directed edges from each variable on the right-hand side of each assignment to the variables on the left-hand side; see Bongers et al. (2021) for a more formal introduction to SCMs.

Let us assume that \(S\) induces a unique observational distribution \(Q\) over \(X^1, \ldots, X^d\) (which is the case if the graph is acyclic, for example), and assume that \(Q\) admits a joint density \(q\) with respect to a product measure. Then \(q\) satisfies the factorisation property (see Lauritzen et al., 1990 or Theorem 1.4.1 in Pearl, 2009): \(q(x^1, \ldots, x^d) = \prod_{j=1}^d q_{X^j|PA^j}(x^j|x^{PA^j})\). In an SCM, an intervention on a variable \(X^k\) replaces the tuple \((f^k, PA^k, Q_{N^k})\) with \((\tilde{f}^k, PA^k, \tilde{Q}_{N^k})\) in the structural assignment for \(X^k\), and we denote the replacement by \(do(X^k := \tilde{f}^k(PA^k, N^k))\) (Pearl, 2009). This new mechanism determines a conditional that we denote by \(P^*(x^k|x^{PA^k})\). The interventional distribution is the induced distribution with the new structural assignment, and we denote this by \(P := Q^{do(X^k := \tilde{f}^k(PA^k, N^k))}\). If \(P\) admits the density \(p\), only the conditional density of \(X^k\) changes

\[\text{for notational convenience we sometimes refer to the parent sets by their indices, i.e., } PA^j \subseteq \{1, \ldots, d\} \setminus \{j\}.\]
(e.g., Aldrich, 1989; Haavelmo, 1944; Pearl, 2009; Peters et al., 2017), that is, for \( j \neq k \), we have 
\[
p(x_j'|x^{PA_j}) = q(x_j'|x^{PA_j}),
\]
for all \( x_j \) and \( x^{PA_j} \). Assume that for the true but unknown distribution \( Q^* \) we know the conditional \( q^*(x_k'|x^{PA_k}) \) (e.g., because this was part of the design when generating the data). Due to the factorisation property, the intervention \( do(X_k: \bar{f}_k(\bar{PA}_k, N_k)) \) can then be represented as a map \( \tau \) that acts on the density \( q \):

\[
\tau(q)(x_1, \ldots, x_d):= p^*(x_k'|x^{PA_k}) \cdot q(x_1, \ldots, x_d).
\]

Defining \( r(x_{\{k\} \cup PA_k \cup \bar{PA}_k}):= p^*(x_k'|x^{PA_k})/q^*(x_k'|x^{PA_k}) \), this takes the form of (5). As the conditional \( p^*(x_k'|x^{PA_k}) \) is fully specified by the intervention, we therefore know the function \( r \). Our proposed framework allows us to test statements about the distribution \( Q^{do(X_k:=\bar{f}_k(\bar{PA}_k, N_k))} \). We obtain similar expressions when intervening on several variables at the same time.

Similar distributional shifts can be obtained, of course, if the factorisation is noncausal (see also Section 3.1), so while our framework contains intervention distributions as a special case, it equally well applies to noncausal models.

**Appendix B: Efficient computation of \( V(n, m) \) in Theorem 4**

In this section, we show that for \( n, m \in \mathbb{N} \) and \( K \geq 1 \)

\[
V(n, m) = \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{\ell} \binom{n-m}{m-\ell} (K^\ell - 1)
\]

can be evaluated efficiently. If \( mn/2 \), such that for some \( \ell \) one has \( m - \ell \geq n - m \), we use the convention that if \( a > b \) then \( \binom{b}{a} = 0 \). If one evaluated the term \( \binom{n}{m}^{-1} \) separately, this could potentially cause numerical underflow, and similarly terms in the sum could get very large, such as the summand including \( K^m - 1 \).

Denote the summands by \( s_{\ell} \), that is

\[
s_{\ell} = \binom{\ell}{l} \frac{(n-m)(m-\ell)}{m-\ell} (K^\ell - 1). \]

We can compute \( s_1 \) by:

\[
s_1 = \binom{m}{1} \frac{(n-m)}{m-1} (K - 1)
\]

\[
= m^2 (K - 1) \frac{(n-m)!}{n!} \frac{(n-m)!}{n-2m+1}!
\]

\[
= (K - 1) m^2 \frac{n^2}{n-m+1} \prod_{j=0}^{m-2} \frac{n-m-j}{n-j}.
\]
This can be evaluated in $O(m)$ time. Further, if $s_\ell \neq 0$, the ratio of two consecutive summands is

$$\frac{s_{\ell+1}}{s_\ell} = \left( \frac{m}{\ell + 1} \right) \left( \frac{n - m}{m - \ell - 1} \right) (K^{\ell+1} - 1) \left( \frac{m}{\ell} \right) \left( \frac{n - m}{m - \ell} \right) (K^\ell - 1) = \frac{(m - \ell)^2}{(\ell + 1)(n - 2m + \ell + 1)} \frac{K^{\ell+1} - 1}{K^\ell - 1},$$

which for a given $\ell$, can be evaluated in $O(1)$ time. Hence, we can compute $\sum_{\ell=1}^m s_\ell$, by first computing $s_1$, and for each $\ell$, compute $s_{\ell+1} = \frac{s_{\ell+1}}{s_\ell} s_\ell$, as long as $s_\ell \neq 0$ (after which the remaining terms are 0). The overall computational cost of computing $V(n, m) = \sum_{\ell=1}^m s_\ell$ is thus $O(m)$.

**B.1 Plotting the level upper bound**

To illustrate how the bound in Theorem 4 depends on $K = \mathbb{E}_Q[r(X)^2]$, $\alpha_\phi$, $m$, and $n$, we plot the level bound $\inf_{\beta \in [0,1]} \left( \frac{\alpha_\phi}{1 - \alpha_\phi} + \frac{V(n, m)}{V(n, m) + \beta} \right)$, for various values of $\alpha_\phi$, $n$, $m$, and $K$ in Figure B.1. By

**Figure B.1.** The finite-level upper bound in Theorem 4 plotted for various values of $m$, $n$, $\alpha_\phi$, and $K$. Here $K = \mathbb{E}_Q[r(X)^2]$ is a measure of the difference between the observed and the target domains and $\alpha_\phi$ is the level of the target test applied.
choosing a very small resample size (such as \( m = n^{0.2} \)) and a conservative target test (such as \( \alpha_p = 0.0001 \)), the level bound grows very slowly in \( K \). For larger resample sizes \( m \), the finite level is only guaranteed when \( K \) is close to 1, meaning that \( P^* \) and \( Q^* \) are similar. As we note in Section 4.4.1, the finite-level bound is in many cases weak, will lead to choices of \( m \) which reduces power, and we therefore in practice recommend to use the GOF-heuristic (Section 4.4.1) in combination with the procedure for combining multiple resamples (Section 4.4.2) instead.

### Appendix C: Algorithm for hypothesis testing with unknown distributional shift

This section contains Algorithm 3, which describes our method for testing under distributional shifts for the case where the shift factor \( r \) is unknown, but can be estimated by an estimator \( \hat{r} \). Algorithm 3 is similar to Algorithm 1 but one additionally splits the sample \( X_n \) into two disjoint samples \( X_{n_1} \) and \( X_{n_2} \) and uses \( X_{n_1} \) for estimating the weights \( \hat{r}_{n_1} \), which are then, together with \( X_{n_2} \), used as an input to Algorithm 1.

We view the sample splitting as a theoretical device. In practice, we are using the full sample both for estimating the weights and for applying the test.

**Algorithm 3**  Testing a target hypothesis with unknown distributional shift and resampling

**Input:** Data \( X_n \), target sample size \( m \), hypothesis test \( \varphi_m \), estimator \( \hat{r} \) for \( r \), and \( a_0 \).

1: Let \( n_1 \), \( n_2 \) be s.t. \( n_1 + n_2 = n \) and \( n_1^2 = \sqrt{n_2} \)
2: \( X_{n_1} \leftarrow X_1, \ldots, X_{n_1} \)
3: \( X_{n_2} \leftarrow X_{n_1+1}, \ldots, X_{n_1+n_2} \)
4: \( \hat{r}_{n_1} \leftarrow \) estimate of \( r \) based on \( X_{n_1} \)
5: \( (i_1, \ldots, i_m) \leftarrow \) sample from \( \{1, \ldots, n_2\}^m \) with weights \( (10) \) on \( \hat{r}_{n_1} \).
6: \( \psi_{\text{DRPL}}^{i_1, \ldots, i_m}(X_{n_2}, U) \leftarrow (X_{n_1+i_1}, \ldots, X_{n_1+i_m}) \)
7: \( \return \psi_{\alpha}(X_{n_2}, U) := \varphi_m(\psi_{\text{DRPL}}^{i_1, \ldots, i_m}(X_{n_2}, U)) \)

### Appendix D: Sampling from \( \Psi_{\text{DRPL}} \)

This section provides details on sampling from \( \Psi_{\text{DRPL}}^{i_1, \ldots, i_m} \) as defined by (10). We have defined \( \Psi_{\text{REPL}}^{i_1, \ldots, i_m} \) and \( \Psi_{\text{NO-REPL}}^{i_1, \ldots, i_m} \) as weighted resampling with and without replacement, respectively. \( \Psi_{\text{REPL}}^{i_1, \ldots, i_m} \) can be implemented as a sequential procedure that first draws \( i_1 \) with weights \( r(X_1)/\sum_{j=1}^n r(X_j) \), and then draws \( i_2 \) with weights \( r(X_i)/\sum_{j=1,j\neq i}^n r(X_j) \), and so forth. Although both \( \Psi_{\text{REPL}}^{i_1, \ldots, i_m} \) and \( \Psi_{\text{NO-REPL}}^{i_1, \ldots, i_m} \) sample distinct sequences \( (i_1, \ldots, i_m) \), they are, in general, not equivalent, as can be seen from the form of the weights \( \psi_{\text{REPL}}^{i_1, \ldots, i_m}(X_{n_1+i_m}, U) \) and \( \psi_{\text{NO-REPL}}^{i_1, \ldots, i_m}(X_{n_1+i_m}, U) \) below. When there is no ambiguity, we omit superscripts and write \( \Psi_{\text{REPL}} \) for example. We also interchangeably consider a sample from \( \Psi_{\text{DRPL}} \) to be a sequence \( (i_1, \ldots, i_m) \) and a subsample \( (X_{i_1}, \ldots, X_{i_m}) \) of \( X_n \).

The procedure \( \Psi_{\text{DRPL}} \), \( \Psi_{\text{REPL}} \), and \( \Psi_{\text{NO-REPL}} \) sample a sequence \( (i_1, \ldots, i_m) \) with weights \( w_{i_1, \ldots, i_m} \) that are, respectively, given by:

\[
\begin{align*}
p_{\text{DRPL}}^{i_1, \ldots, i_m} &= \frac{\prod_{j=1}^m r(X_{i_j})}{\sum_{i_1, \ldots, i_m} \prod_{j=1}^m r(X_{i_j})} \quad \text{fordistinct } (i_1, \ldots, i_m) \\
p_{\text{REPL}}^{i_1, \ldots, i_m} &= \frac{\prod_{j=1}^m r(X_{i_j})}{\sum_{i_1, \ldots, i_m} \prod_{j=1}^m r(X_{i_j})} \quad \text{forall } (i_1, \ldots, i_m) \\
p_{\text{NO-REPL}}^{i_1, \ldots, i_m} &= \sum_{j_1=1}^n r(X_{j_1}) \sum_{j_2=1}^n r(X_{j_2}) \cdots \sum_{j_m=1}^n r(X_{j_m}) \quad \text{fordistinct } (i_1, \ldots, i_m)
\end{align*}
\]

Here, the comment ‘fordistinct \( (i_1, \ldots, i_m) \)’ implies that the weights are zero otherwise. Most statistical software have standard implementations for sampling from \( \Psi_{\text{REPL}} \) and \( \Psi_{\text{NO-REPL}} \) (known simply as sampling with or without replacement). We now detail a number of ways to sample a
sequence \((i_1, \ldots, i_m)\) from \(\Psi_{\text{DRPL}}\). The first two sampling methods are exact, the third sampling method is approximate.

### D.1 Acceptance–rejection sampling with \(\Psi_{\text{REPL}}\) as proposal

Given a sample \(X_n\), one can sample from \(\Psi_{\text{DRPL}}\) by acceptance–rejection sampling from \(\Psi_{\text{REPL}}\), by drawing sequences \((i_1, \ldots, i_m)\) from \(\Psi_{\text{REPL}}\) until one gets a draw that is distinct, which is then used as the draw from \(\Psi_{\text{DRPL}}\). This is a valid sampling method for \(\Psi_{\text{DRPL}}\), because for any distinct sequence \((i_1, \ldots, i_m)\) we have

\[
P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m) \mid \Psi_{\text{REPL}} \text{ distinct}, X_n) = \frac{P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m), \Psi_{\text{REPL}} \text{ distinct} \mid X_n)}{P_Q(\Psi_{\text{REPL}} \text{ distinct} \mid X_n)}
= \frac{P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m) \mid X_n)}{P_Q(\Psi_{\text{REPL}} \text{ distinct} \mid X_n)}
= \frac{P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m) \mid X_n)}{\sum_{(i_1, \ldots, i_m)} \Psi_{\text{REPL}}(i_1, \ldots, i_m)}
= \frac{\prod_{i=1}^{m} r(X_i) \prod_{i=1}^{m} \Psi_{\text{REPL}}(i_1, \ldots, i_m)}{\sum_{(i_1, \ldots, i_m)} \Psi_{\text{DRPL}}(i_1, \ldots, i_m)}
= \frac{P_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m) \mid X_n)}{\sum_{(i_1, \ldots, i_m)} \Psi_{\text{DRPL}}(i_1, \ldots, i_m)}
= \frac{P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m) \mid X_n)}{\sum_{(i_1, \ldots, i_m)} \Psi_{\text{DRPL}}(i_1, \ldots, i_m)}.
\]

By integrating over \(X_n\), this implies that \(P_Q(\Psi_{\text{REPL}} = (i_1, \ldots, i_m) \mid \Psi_{\text{REPL}} \text{ distinct}) = P_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m))\). Proposition 2 shows that under certain assumptions, the probability of sampling a distinct sample from \(\Psi_{\text{REPL}}\) converges to 1.

### D.2 Acceptance–rejection sampling with \(\Psi_{\text{ND–REPL}}\) as proposal

If \(m\) is large compared to \(n\), it may be that most of the samples drawn from \(\Psi_{\text{REPL}}\) are not distinct, and so the acceptance–rejection scheme in Appendix D.1 may take too many attempts to produce a distinct sample. As an alternative, one can use \(\Psi_{\text{ND–REPL}}\) as a proposal distribution for an acceptance–rejection sampler, which is typically faster, since \(\Psi_{\text{ND–REPL}}\) has the same support as \(\Psi_{\text{DRPL}}\). Given a sample \(X_n\), we thus need to identify an \(M\) such that

\[
\forall \text{ distinct}(i_1, \ldots, i_m): \frac{P_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m) \mid X_n)}{P_Q(\Psi_{\text{ND–REPL}} = (i_1, \ldots, i_m) \mid X_n)} \leq M.
\]

We have

\[
\frac{P_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m) \mid X_n)}{P_Q(\Psi_{\text{ND–REPL}} = (i_1, \ldots, i_m) \mid X_n)} = \frac{\Psi_{\text{DRPL}}(i_1, \ldots, i_m)}{\Psi_{\text{ND–REPL}}(i_1, \ldots, i_m)} = \frac{\prod_{i=1}^{m} r(X_i) \prod_{i=1}^{m} \Psi_{\text{DRPL}}(i_1, \ldots, i_m)}{\sum_{(i_1, \ldots, i_m)} \Psi_{\text{DRPL}}(i_1, \ldots, i_m)}.
\]

The denominator does not depend on \(i_1, \ldots, i_m\), and the numerator can be upper bounded by

\[
(1 - 0)(1 - p(1))(1 - p(1) - p(2)) \cdots (1 - p(1) - \cdots - p(m-1)),
\]

where \(p(1) = \min(r(X_1), \ldots, r(X_n))\) is the smallest of the weights, \(p(2)\) is the second smallest, etc. Thus, we can choose

\[
M := \frac{(1 - 0)(1 - p(1))(1 - p(1) - p(2)) \cdots (1 - p(1) - \cdots - p(m-1))}{\sum_{(i_1, \ldots, i_m)} \prod_{i=1}^{m} r(X_i)}.
\]
We now proceed with an ordinary acceptance–rejection sampling scheme: We sample a (distinct) sequence \((i_1, \ldots, i_m)\) from \(\Psi_{\text{NO–REPL}}\) and an independent, uniform variable \(V\) on the interval \((0, 1)\). We accept \((i_1, \ldots, i_m)\) if

\[
V \leq \frac{\mathbb{P}_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m) \mid X_n)}{M \cdot \mathbb{P}_Q(\Psi_{\text{NO–REPL}} = (i_1, \ldots, i_m) \mid X_n)} = \frac{(1 - 0)(1 - p_{i_1})(1 - p_{i_1} - p_{i_2}) \cdots (1 - p_{i_1} - \cdots - p_{i_{m-1}})}{(1 - 0)(1 - p_{(1)})(1 - p_{(2)} - \cdots (1 - p_{(1)} - \cdots - p_{(m-1)})}.
\]

Here, we have used that the denominator of \(M\) cancels with the normalisation constant of \(\mathbb{P}_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_m) \mid X_n)\). If the sample is not accepted, we draw another sample from \(\Psi_{\text{NO–REPL}}\) until one sample is accepted.

### D.3 Approximate Gibbs sampling starting from \(\Psi_{\text{NO–REPL}}\)

There are cases, where the sampling schemes presented in Appendices D.1 and D.2 do not yield an accepted sample in a reasonable amount of time (this is typically due to \(m\) being too large compared to \(n\), see (A1)). In such cases, one can get an approximate sample of \(\Psi_{\text{DRPL}}\) by sampling \(\Psi_{\text{NO–REPL}}\) and shifting it towards \(\Psi_{\text{DRPL}}\) using a Gibbs sampler (Geman & Geman, 1984).

Let therefore \((i_1, \ldots, i_m)\) be an initial (distinct) sample from \(\Psi_{\text{NO–REPL}}\), and define \(i_{-\ell}\) to be the sequence without the \(\ell\)th entry. The Gibbs sampler sequentially samples \(i_\ell\) from the conditional distribution \(j \mid i_{-\ell}\) in \(\Psi_{\text{DRPL}}\). To compute this conditional probability let \(\Psi_{\text{DRPL}}^\ell\) be the \(\ell\)th index of a sample. Then

\[
\mathbb{P}_Q(\Psi_{\text{DRPL}}^\ell = j \mid \Psi_{\text{DRPL}} = i_{-\ell}) = \frac{\mathbb{P}_Q(\Psi_{\text{DRPL}} = (i_1, \ldots, i_\ell, \ldots, i_m))}{\mathbb{P}_Q(\Psi_{\text{DRPL}} = i_{-\ell})} = \frac{r(X_{i_1}) \cdots r(X_{i_\ell}) \cdots r(X_{i_m})}{\sum_{\nu \neq i_\ell} r(X_{i_1}) \cdots r(X_{i_\nu}) \cdots r(X_{i_m})} = \frac{r(X_1)}{\sum_{\nu \neq i_\ell} r(X_{i_\nu})},
\]

i.e., the conditional distribution of one index \(i_\ell\) given \(i_{-\ell}\) is just a weighted draw among \([i_1, \ldots, i_m]\)\(i_{-\ell}\). This is simple to sample from and the Gibbs sampler now iterates through the indices \((i_1, \ldots, i_m)\), at each iteration replacing the index \(i_\ell\) by a sample from the conditional given \(i_{-\ell}\). Iterating this a large number of times produces an approximate sample from \(\Psi_{\text{DRPL}}\).

### Appendix E: Sampling with replacement

Instead of the sampling scheme \(\Psi_{\text{DRPL}}\) presented above, we can also use weighted sampling with replacement, which we denote \(\Psi_{\text{REPL}}\) (see Appendix D for details). Sampling from \(\Psi_{\text{REPL}}\) is simpler than from \(\Psi_{\text{DRPL}}\), and while sampling from \(\Psi_{\text{REPL}}\) is in some cases disadvantageous for testing (e.g., if we test whether the target distribution has a point mass), if the test is not prone to duplicate data points, testing based on \(\Psi_{\text{REPL}}\) may be advantageous over \(\Psi_{\text{DRPL}}\) (further examination is needed to clarify this relationship). When sampling without weights, Bickel et al. (2012) present regularity conditions on the test statistic that guarantee consistency even with \(m = o(n)\).

Here we show that under additional assumptions, the probability of a nondistinct sample from \(\Psi_{\text{REPL}}\) converges to 0. Consider the following strengthening of (A2).

(A3) There exists \(L \in \mathbb{R}\) such that for all \(v \geq 1\), \(\mathbb{E}_Q[r(X_i)^{v+1}] \leq L^v\).

This is for instance trivially satisfied if \(r(X_i)\) is \(Q\)-a.s. bounded by a constant \(L\). The following proposition shows that under (A1) and (A3) the probability of drawing a distinct sample from \(\Psi_{\text{REPL}}\) converges to 1.
Proposition 2  (Asymptotic equivalence of REPL and DREPL for bounded weights). Let $\tau: \mathcal{Q} \rightarrow \mathcal{P}$ be a distributional shift for which a known map $r: \mathcal{X} \rightarrow [0, \infty)$ exists, satisfying $r(q)(x) \propto r(x)q(x)$, see (5). Consider an arbitrary $Q \in \mathcal{Q}$ and $P = \tau(Q)$. Let $m = m(n)$ be a resampling size and let $\Psi_{\text{REPL}}$ be the weighted resampling with replacement defined in Section 4.1. Then, if $m$ and $Q$ satisfy (A1) and (A3), it holds that

$$\lim_{n \to \infty} P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ distinct}) = 1.$$ 

As a corollary to Theorem 1, we also have pointwise asymptotic level of a test when $\Psi_{\text{REPL}}$ is used instead of $\Psi_{\text{DREPL}}$.

Corollary 3  (Pointwise asymptotics—REPL). Assume the same setup and assumptions as in Theorem 1 and additionally assume (A3). Let $\Psi_{\text{REPL}}$ be the weighted resampling with replacement defined in Section 4.1 and let $\psi_n^r$ be the REPL-based resampling test defined by $\psi_n^r(X_n, U) := \varphi_m(\Psi_{\text{REPL}}^m(X_n, U))$. Then, it holds that

$$\limsup_{n \to \infty} P_Q(\psi_n^r(X_n, U) = 1) = \alpha \varphi.$$ 

The same statement holds when replacing both $\limsup$’s (including the one in $\alpha \varphi$) with $\liminf$’s.

Appendix F: Algorithm for the GOF-heuristic for choosing $m$

The GOF-heuristic is a data driven procedure to choose $m$ in finite sample settings, see Section 4.4.1. It is summarised in Algorithm 4.

Algorithm 4  GOF-heuristic: Choosing $m$ by testing resampling validity

Input: Data $X_n$, shift factor $r(x^A)$, threshold $\alpha_c$, initial target size $m_0$, increment size $\Delta$, repetitions $K$, conditional goodness-of-fit test $\kappa$.

1: $q_\alpha \leftarrow \alpha_c$-quantile of $\text{mean}(U_1, \ldots, U_K)$, where $U_i \sim \text{Unif}(0, 1)$
2: $m \leftarrow m_0$
3: $m_{\text{valid}} \leftarrow \text{true}$
4: while $m_{\text{valid}}$ do
5:     for $k = 1, \ldots, K$ do
6:         $\text{res}_k \leftarrow \kappa(\Psi_{\text{REPL}}^m(X_n))$
7:     if $\text{mean}(\text{res}_1, \ldots, \text{res}_K) > q_\alpha$ then
8:         $m \leftarrow m + \Delta$
9:     else
10:         $m_{\text{valid}} \leftarrow \text{false}$
11:     $m \leftarrow m - \Delta$
12: return $m$

Appendix G: Additional experiments

G.1 Assumption (A1) when sample size increases

As indicated by the results in Section 5.1, (A1) may sometimes be too strict of an assumption, in the sense that level can be attained also when (A1) is violated. We explored this for a continuous data example in Section 5.1, but here, we investigate the effect of violating (A1) as the sample size $n$ increases for a binary setting.
We simulate data $X_n$ from a binary distribution $P(X = 0) = 0.9$, $P(X = 1) = 0.1$, and consider the target distribution $P$ where the probabilities are flipped, i.e., $P(X = 0) = 0.1$, $P(X = 1) = 0.9$. We consider the hypothesis $H_0 : P_P(X = 1) > 0.8$, which is clearly satisfied for $P$, but may not be detected by the resample if $m = n^a$ for $a > 0.5$ chosen too large. We repeat the experiment 200 times and compute the resulting rejection rates for various sample sizes and rates $m = n^a$. The results are shown in Figure G.1. Under (A1), Theorem 1 guarantees asymptotic level. Indeed, at all sample sizes $n$, the test has the correct level, when (A1) is satisfied (that is, when $a < 0.5$). Yet, though this is not guaranteed by theory, we have indications of asymptotic level also for $a > 0.5$.

G.2 Model selection under covariate shift

In this section, we apply our testing method to the problem of model selection under covariate shift as discussed in Section 3.6. We generate a data set $D := \{(X_i, Y_i)\}_{i=1}^n$ of size $n = 3'000$. Each $(X_i, Y_i)$ is drawn i.i.d. according to the following data generating process:

$$(X^1, X^2) \sim \text{GaussianMixture}([3, 3], [−3, −3]) \quad Y_i := \begin{cases} \sin(X_2 + \epsilon_y), & \text{if } X_1 \geq 0 \\ \sin(3 + X_1 + \epsilon_y), & \text{if } X_1 < 0, \end{cases}$$

where $\text{GaussianMixture}([3, 3], [−3, −3])$ is an even mixture (i.e., $p = 0.5$) of two 2-dimensional Gaussian distributions with means $\mu_1 = (3, 3)^\top$, $\mu_2 = (−3, −3)^\top$ and unit covariance matrix, and $\epsilon_y$ is a standard Gaussian $\mathcal{N}(0, 1)$-variable. Figure G.2 illustrates a sample of size $1'000$ from this data generating process.

We randomly split the data $D$ into a training set $D_{\text{train}}$ of size $2'000$ and a test set $D_{\text{test}}$ of size $1'000$. Using $D_{\text{train}}$, we train two candidate classifiers, namely logistic regression (LR) and random forest (RF) to predict $Y$ from $X$. Both models are trained using the Scikit-Learn Python package (Pedregosa et al., 2011) with default parameters. We consider the area under the curve (AUC) as the scoring function, where we denote the AUC scores for the models LR and RF by $\text{AUC(LR)}$ and $\text{AUC(RF)}$, respectively. Then, we apply our resampling approach on $D_{\text{test}}$ to test whether LR outperforms RF when the distribution of $(X^1, X^2)$ is changed to a single 2-dimensional Gaussian distribution with mean $\mu = (3, 3)^\top$ and unit covariance matrix. In this experiment, we choose the resampling size $m$ by the GOF-heuristic (see Algorithm 4) and assume that the shift factor $r(x) := p^*(x)/q^*(x)$ is known, where $q^*(x)$ is a pdf of the mixture of Gaussian $\text{GaussianMixture}([3, 3], [−3, −3])$ and $p^*(x)$ is a pdf of the Gaussian $\mathcal{N}((3, 3)^\top, I_2)$. We employ DeLong’s test (DeLong et al., 1988) to test the hypothesis $\text{AUC(LR)} \leq \text{AUC(RF)}$ using the R-package pROC (Robin et al., 2011).

We repeat the experiment 500 times and report in Table G.1 how many times our resampling test (resampling) rejects and returns that $\text{AUC(LR)} > \text{AUC(RF)}$. As a benchmark, we also report the corresponding rejection rate when we perform the test directly on a sample from the target distribution (oracle) in which the covariate distribution is changed to $p^*(x)$. We also report the rejection rate when we perform the test on the observed test set directly $D_{\text{test}}$ (observed), without resampling it first.

As shown in Table G.1, under the observed distribution we have $\text{AUC(LR)} \leq \text{AUC(RF)}$ (the rejection rate is 0 in the observed sample). However, under the target distribution $\text{AUC(LR)}$ is higher than $\text{AUC(RF)}$ (the rejection rate is 1 in the target sample). Our resampling approach yields high power against the alternative hypothesis, even without having access to the oracle information but using resampling instead.

G.3 Conditional independence testing in the bikeshare dataset

Following Berrett et al. (2020), we consider the capital bikeshare dataset (Fanaee-T & Gama, 2014). Berrett et al. (2020) test conditional independence of the duration $X$ from three different outcomes $Y_j$, $j = 1, 2, 3$, where $Y_1$ is a binary ‘User type’ variable indicating whether the user is a member, $Y_2$ is the date and $Y_3$ is the day of the week, conditioning on a variable $Z$ which encodes mean and variance for the particular ride and at that time of the day. More specifically, they use their proposed permutation test (CPT) and the randomisation test (CRT) by Candès et al. (2018). Here, we apply our method for testing conditional independence from Section 3.1 to the same data, using the GOF-heuristic (Section 4.4.1) to chose $m$. Since both ours and their method rely
on knowing the conditional \( q(x|z) \), we can use their estimated conditionals \( \hat{q}(x|z) \) such that differences in outcome is not because of differences in estimation.

We display the resulting \( p \)-values for the hypothesis of conditional independence in Table G.2. At a 5\% significance level, our method rejects the same hypotheses as they do, finding that the duration is not conditionally independent of the user type, but is so of both the date and the day of the week.

G.4 Testing dormant independence in a nonlinear, nonGaussian model

In this appendix, we consider an experiment similar to those in Section 5.6, but with a nonlinear SCM with nonlinear errors (see details in Appendix G.5).

We estimate the conditional \( q(x^3|x^2) \) using generalised additive models. For simplicity, we consider a distribution where \( X^3 - \mathbb{E}[X^3|X^2] \) is Gaussian, though our procedure also applies to more general settings, for example by using a generalised linear model or by applying conditional density estimation to learn \( q(x^3|x^2) \).

We plot the resulting rejection rates in Figure G.3. We observe that also in this setting, our method has the correct level and increasing power as sample size increases. This highlights that our method is not restricted to binary or Gaussian data, but applies to all models where \( q(x^3|x^2) \) is known or can be estimated. Since the direct effect of \( X^1 \) of \( X^4 \) is a shift in the mean of \( X^4 \) (see Appendix G.5), one could in principle also apply the methodology of Robins (1999) to this problem, though we did not find any reasonable implementation of this.

G.5 Additional simulation details

In this section, we state the models used for generating the data in Sections 5.2, 5.6, 5.9 and 5.10 and Appendix G.4.
Section 5.2

The linear data in Section 5.2 was generated by the equations

\[ Z := \varepsilon_Z \quad X := Z + 2\varepsilon_X \quad Y := Z + \theta X + 2\varepsilon_Y \]

and the nonlinear data was generated by the equations

\[ X := \varepsilon_X \quad Z := |X - 1| + 4\varepsilon_Z \quad Y := Z + \theta X + 2\varepsilon_Y \]

where \( \varepsilon_Z, \varepsilon_X, \varepsilon_Y \sim \mathcal{N}(0, 1) \), and either \( \theta = 0 \) (no effect present) or \( \theta = 0.5 \) (effect present).

Section 5.6 and Appendix G.4

The binary data in Section 5.6 was generated by first sampling hyper-parameters:

\[ p_H \sim \text{Dirichlet}^{1\times 4}(3, 3, 3, 3) \quad p_1 \sim B^{1\times 1}(1, 1) \quad p_2 \sim B^{2\times 4}(1, 1)p_3 \sim B^{1\times 2}(1, 1) \]

\[ G : p_4 \sim B^{2\times 1}(1, 1) \quad H : p_4 \sim B^{2\times 2}(1, 1), \]

Table G.1. Rejection rates of the test with alternative AUC(LR) > AUC(RF) over 500 repetitions. The resampling test shows high power to detect the alternative

<table>
<thead>
<tr>
<th>Resampling</th>
<th>Oracle</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.874 ± 0.029</td>
<td>1.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>

Figure G.2. Sample from the data generating process for the experiment described in Appendix G.2.

Table G.2. Comparison of \( p \)-values for the conditional independence hypothesis \( X \perp Y_j | Z \) for three different variables \( Y_j \). We compare to CPT (Berrett et al., 2020) and CRT (Candès et al., 2018) to our method using a single test or the combination test of Hartung (1999) as discussed in Section 4.4.2

<table>
<thead>
<tr>
<th>Variable</th>
<th>CPT</th>
<th>CRT</th>
<th>Ours (combination test)</th>
<th>Our (single test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>User type</td>
<td>0.0010</td>
<td>0.0010</td>
<td>1.447 \cdot 10^{-16}</td>
<td>4.595 \cdot 10^{-5}</td>
</tr>
<tr>
<td>Date</td>
<td>0.1146</td>
<td>0.1293</td>
<td>0.1320</td>
<td>0.0586</td>
</tr>
<tr>
<td>Day of week</td>
<td>0.1980</td>
<td>0.2063</td>
<td>0.4395</td>
<td>0.1246</td>
</tr>
</tbody>
</table>
where \(B(a, b)\) is a Beta distribution and the superscript indicates the dimension of the sampled parameter matrix. The matrices \(p_1, \ldots, p_4\) correspond to conditional probability tables given parent variables in the graphs \(G\) or \(H\). The distributions are the same when sampling from \(G\) and \(H\), except for \(X^4\), which has an additional parent in \(H\). In each repetition, given hyper-parameters, a data set is sampled from the structural equation model

\[
G: X^4 \sim \text{Bernoulli}(p_4) \quad H: X^4 \sim \text{Bernoulli}(p_{4, X^1})
\]

where the subscript \(p_{2, X^1, U}\) indicate that for an outcome of \((X^1, U)\), the Bernoulli distribution uses the probability in the corresponding entry of \(p_2\) (and similar for \(p_3\) and \(p_4\)).

The Gaussian data in Section 5.6 was generated by the structural equation model

\[
H := e_H \quad X^1 := e_{X^1} \quad X^2 := X^1 + H + e_{X^2} \\
X^3 := X^2 + 2e_{X^3} \quad X^4 := \theta \cdot X^1 + X^3 + H + e_{X^4}
\]

where \(e_H, e_{X^1} \sim \mathcal{N}(0, 1)\), and \(\theta \in \{0, 0.3\}\) indicates the absence or presence of the edge \(X^1 \rightarrow X^4\).

The nonGaussian data in Appendix G.4 was generated by the structural equation model

\[
H := \frac{1}{2} \cdot e_H \cdot e_{H} \quad X^1 := \gamma_{X^1} \quad X^2 := X^1 \cdot H + e_{X^2} \\
X^3 := X^2 \cdot X^2 + \frac{3}{2} e_{X^3} \quad X^4 := \theta \cdot X^1 + X^3 + H + e_{X^4},
\]

where \(e_H, e_{X^1} \sim \mathcal{N}(0, 1)\) and \(\gamma_{X^1}\) follows a \(\Gamma(2)\)-distribution, and \(\theta \in \{0, 0.3\}\) indicates the absence or the presence of the edge \(X^1 \rightarrow X^4\).

**Section 5.9**

The data in Section 5.9 was generated by the structural equation model

\[
X^1 := e_{X^1} \quad X^3 := e_{X^3} \quad X^2 := X^3 + X^1 + 2e_{X^1} \\
Y := X^2 + X^3 + 0.3e_Y \quad X^4 := -Y + X^2 + X^3 + 0.7e_{X^4}
\]
**Section 5.10**
The linear data in Section 5.10 was generated by the equations

\[ X := \varepsilon_X \quad \mathbb{P}(D = 1) = \sigma \left( \frac{2}{\sqrt{20}} \sum_{i=1}^{20} X_i \right) \]

and the nonlinear data was generated by the equations

\[ X := \varepsilon_X \quad \mathbb{P}(D = 1) = \sigma \left( \frac{2}{\sqrt{20}} \sum_{i=1}^{20} X_i \right) \]

\[ Y := \theta D + (1 + \theta D) \left( 0.2 \sum_{i=1}^{20} X_i + \exp \left( -0.2 \left( \sum_{i=1}^{20} X_i \right)^2 / 2 \right) \right) \sin \left( 0.2 \sum_{i=1}^{20} X_i \right) + 0.3 \varepsilon_Y \]

where \( \varepsilon_X \sim \mathcal{N}(0, \text{Id}_{20}) \) and \( \varepsilon_Y \sim \mathcal{N}(0, 1) \), and either \( \theta = 0 \) (no effect present) or \( \theta = 0.3 \) (effect present).

**Appendix H: Proofs**

**H.1 Proof of Theorem 1**

**Proof of Theorem 1.** We show the statement only for lim sup, the corresponding statement for lim inf follows by replacing lim sup with lim inf everywhere.

Let \( p \) and \( q \) denote the respective densities of \( P \) and \( Q \) with respect to the dominating measure \( \mu \). By assumption \( p = r(q) \), so \( p(x) \propto r(x)q(x) \). Let \( \bar{r} \) be the normalised version of \( r \) satisfying \( p(x) = \bar{r}(x)q(x) \). Recall that we call a sequence \( (i_1, \ldots, i_m) \) distinct if for all \( \ell \neq \ell' \) we have \( i_\ell \neq i_{\ell'} \). The resampling scheme \( \Psi_{\text{DRPL}} \), defined by (10), samples from the space of distinct sequences \( (i_1, \ldots, i_m) \), where every sequence has probability \( w_{i_1,\ldots,i_m} \propto \prod_{\ell=1}^{m} r(X_{i_\ell}) \). The normalisation constant here is the sum over the weights in the entire space of distinct sequences, that is,

\[
\omega_{i_1,\ldots,i_m} = \frac{\prod_{\ell=1}^{m} r(X_{i_\ell})}{\sum_{\ell_1,\ldots,\ell_m \text{ distinct}} \prod_{\ell=1}^{m} r(X_{i_\ell})}.
\]

Thus, taking an expectation involving \( \varphi_m(\Psi_{\text{DRPL}}^{\text{em}}(X, U)) \), amounts to evaluating \( \varphi_m \) in all distinct sequences \( X_{i_1}, \ldots, X_{i_m} \) and weighting with the probabilities \( w_{i_1,\ldots,i_m} \).

\[
E_Q \left[ \frac{1}{n!} \sum_{i_1,\ldots,i_m \text{ distinct}} (\prod_{\ell=1}^{m} r(X_{i_\ell})) 1_{\varphi_m(X_{i_1,\ldots,i_m}) = 1} \right] = \frac{1}{n!} \sum_{i_1,\ldots,i_m \text{ distinct}} (\prod_{\ell=1}^{m} r(X_{i_\ell}))
\]

where we divide by the number of distinct sequences \( \frac{n!}{(n-m)!} \) in both numerator and denominator.
Let $c(n, m)$ and $d(n, m)$ be the numerator and denominator terms of (H1), i.e.,

$$c(n, m) := \frac{1}{n!} \sum_{(i_j: j \neq 0) \text{ distinct}} \left( \prod_{i=1}^{m} \bar{r}(X_{i_j}) \right) \mathbb{1}_{(\varphi_m(X_{i_1}, \ldots, X_{i_m}) = 1)},$$

$$d(n, m) := \frac{1}{(n - m)!} \sum_{i \in \{0, \ldots, n - m\}} \prod_{i=1}^{m} \bar{r}(X_{i_j}).$$

We want to show that $\limsup_{n \to \infty} \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = 0$. To see this, define for all $\delta > 0$ the set $A_\delta := \{|d(n, m) - 1| \leq \delta\}$. It holds for all $\delta \in (0, 1)$ that

$$\mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{A_\delta} \right] + \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{\overline{A_\delta}} \right]$$

$$\leq \mathbb{E}_Q \left[ \frac{c(n, m)}{1 - \delta} \mathbb{1}_{A_\delta} \right] + \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{\overline{A_\delta}} \right]$$

$$\leq \mathbb{E}_Q \left[ \frac{c(n, m)}{1 - \delta} \right] + \mathbb{P}_Q(A_{\delta})$$

$$= \frac{1}{1 - \delta} \mathbb{P}_Q(\varphi_m(X_1, \ldots, X_m) = 1) + \mathbb{P}_Q(A_{\delta}),$$

where we used that $\frac{c(n, m)}{d(n, m)} \leq 1$ and Lemma 1 (a). Further combining Chebyshev’s inequality with Lemma 1 (b) and (d), it follows that

$$\lim_{n \to \infty} \mathbb{P}_Q(A_{\delta}^c) \leq \lim_{n \to \infty} \frac{\mathbb{E}_Q[(d(n, m) - 1)^2]}{\delta^2} = \lim_{n \to \infty} \frac{\mathbb{V}_Q(d(n, m))}{\delta^2} = 0.$$

Hence, using that $\limsup_{k \to \infty} \mathbb{P}_P(\varphi_k(X_1, \ldots, X_k) = 1) = \alpha_P$ we have shown for all $\delta \in (0, 1)$ that

$$\limsup_{n \to \infty} \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] \leq \frac{1}{1 - \delta} \alpha_P. \quad (H2)$$

Similarly, we also get for all $\delta \in (0, 1)$ the following lower bound:

$$\mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{A_\delta} \right] + \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{\overline{A_\delta}} \right]$$

$$\geq \mathbb{E}_Q \left[ \frac{c(n, m)}{1 + \delta} \mathbb{1}_{A_\delta} \right]$$

$$\geq \frac{1}{1 + \delta} \mathbb{E}_Q[c(n, m)],$$

where in the last inequality we used that $c(n, m) \geq 0$. Again using Lemma 1 (a) and that $\limsup_{k \to \infty} \mathbb{P}_P(\varphi_k(X_1, \ldots, X_k) = 1) = \alpha_P$, we get that for all $\delta \in (0, 1)$ that

$$\limsup_{n \to \infty} \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] \geq \frac{1}{1 + \delta} \alpha_P. \quad (H3)$$

Combining both limits (H2) and (H3) and using that $\delta \in (0, 1)$ is arbitrary, this proves that

$$\limsup_{n \to \infty} \mathbb{E}_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = \alpha_P,$$

which completes the proof of Theorem 1. □
Remark 1  After Theorem 2, we discussed the conjecture that for any $\ell \geq 2$ we can relax (A1) to $m = o(n^{1-1/\ell})$ if one changes (A2) to $\mathbb{E}_Q[r(X_i)^\ell]<\infty$. We now argue why the current structure of the proof of Theorem 1 does not allow for proving this statement. In the proof, we use that $\forall Q[d(n,m)]$ converges to 0, and the expression for the variance for U-statistics that we use in (H7) in Lemma 1 only depends on the second moment of the weights. In particular, the vanishing of the variances only depends on the relation between $m$ and $n$, and not higher order moments of $\mathbb{E}[r(X_i)^\ell]$, and even if those were finite, the counterexample in Theorem 2 shows that the variance of $c(n, m)$ and $d(n, m)$ would still approach infinity if $m = o(\sqrt{n})$, i.e., $m$ grows at least as fast as $\sqrt{n}$.

Thus, if one were to prove a result for the case $\mathbb{E}_Q[r(X_i)^\ell]<\infty$, it appears to us that one needs to change the proof of Theorem 1 to treat $\mathbb{E}_Q[c(n, m)/d(n, m)]$ jointly rather than using Lemma 1 to treat each of $c(n, m)$ and $d(n, m)$ separately.

Lemma 1  (Distinct draws). Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ have densities $p$ and $q$ with respect to a dominating measure $\mu$. Let $\bar{r}: \mathcal{X} \to [0, \infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) = \bar{r}(x)q(x)$. Let $c(n, m)$ and $d(n, m)$ be defined by

\[
c(n, m) := \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \left( \prod_{i=1}^{m} \bar{r}(x_i) \right) \delta_{\{p(x_1, \ldots, x_m) = 1\}},
\]

\[
d(n, m) := \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \prod_{i=1}^{m} \bar{r}(x_i).
\]

Then, if $m$ and $Q$ satisfy (A1) and (A2) it holds that

(a) $\mathbb{E}_Q[c(n, m)] = \mathbb{P}_P[q_{m}(X_1, \ldots, X_m) = 1]$,
(b) $\mathbb{E}_Q[d(n, m)] = 1$,
(c) $\lim_{n \to \infty} \mathbb{V}_Q[c(n, m)] = 0$,
(d) $\lim_{n \to \infty} \mathbb{V}_Q[d(n, m)] = 0$.

Proof. (A) we first prove the statements for the means, i.e., (a) and (b), and (B) we then prove the statements for the variances, i.e., (c) and (d).

Part A (means): Define $\delta_m := \delta_{\{p(x_1, \ldots, x_m) = 1\}}$ (for the case (H4)) or $\delta_m := 1$ (for the case (H5)). Then, in both cases it holds that

\[
\mathbb{E}_Q\left[ \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \left( \prod_{i=1}^{m} \bar{r}(x_i) \right) \delta_m \right]
\]

\[
= \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \mathbb{E}_Q \left[ \prod_{i=1}^{m} \bar{r}(x_i) \right] \delta_m
\]

\[
= \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \int \left( \prod_{i=1}^{m} \bar{r}(x_i)q(x_i) \right) \delta_m p_{\mu}(x_1, \ldots, x_m)
\]

\[
= \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \int \left( \prod_{i=1}^{m} \bar{r}(x_i) \right) \delta_m p_{\mu}(x_1, \ldots, x_m)
\]

\[
= \frac{1}{n!} \sum_{\{x_1, \ldots, x_m\} \text{ distinct}} \mathbb{E}_P[\delta_m]
\]

\[
= \mathbb{E}_P[\delta_m]
\]
In the second and fourth equality, we use that \(i_1, \ldots, i_m\) are all distinct, and in the last equality, we use that the number of distinct sequences \((i_1, \ldots, i_m)\) is \(\frac{n!}{(n-m)!}\). Consequently, the term in \((H4)\) has mean \[\mathbb{E}[\varphi_m(X_{i_1}, \ldots, X_{i_m}) = 1] = \mathbb{P}(\varphi_m(X_1, \ldots, X_m) = 1)\] and the term in \((H5)\) has mean 1.

**Part B (variances):** We begin by expressing \[\frac{1}{n!} \sum_{(i_1, \ldots, i_m) \text{ distinct}} h_m(Z_{i_1}, \ldots, Z_{i_m})\] as a U-statistic (Serfling, 1980). A U-statistic has the form

\[\frac{1}{n!} \sum_{(i_1, \ldots, i_m) \text{ distinct}} \left( \prod_{\ell=1}^m \bar{r}(X_{i_{\ell}}) \right) \delta_m\] for some symmetric function \(h_m(z_1, \ldots, z_m)\) (called a kernel function). In our case, the kernel function is \(h_m(X_{i_1}, \ldots, X_{i_m}) := \prod_{\ell=1}^m \bar{r}(X_{i_{\ell}})\delta_m\). The variance of the corresponding U-statistic (see Serfling, 1980, Section 5.2) is given by

\[\mathbb{V}_Q \left( \frac{1}{n!} \sum_{(i_1, \ldots, i_m) \text{ distinct}} \left( \prod_{\ell=1}^m \bar{r}(X_{i_{\ell}}) \right) \delta_m \right) = \left( \frac{n}{m} \right)^{-1} \sum_{v=1}^m \left( m \atop v \right) \left( n - m \right) \left( m - v \right) \zeta_v\] (H7)

where for all \(v \in \{1, \ldots, m\}\)

\[\zeta_v := \mathbb{E}_Q[\mathbb{E}_Q[h_m(X_{i_1}, \ldots, X_{i_m}) \mid X_{i_1}, \ldots, X_{i_v}]].\]

We now bound \(\zeta_v\) from above by the second moment as follows:

\[\zeta_v \leq \mathbb{E}_Q[\mathbb{E}_Q[h_m(X_{i_1}, \ldots, X_{i_m}) \mid X_{i_1}, \ldots, X_{i_v}]^2].\]

Moreover, using that \(\delta_m\) is upper bounded by 1, we get for both cases \((H5)\) and \((H4)\) that

\[\zeta_v \leq \mathbb{E}_Q[\mathbb{E}_Q[h_m(X_{i_1}, \ldots, X_{i_m}) \mid X_{i_1}, \ldots, X_{i_v}]^2]\]

\[\leq \mathbb{E}_Q \left[ \mathbb{E}_Q \left[ \prod_{\ell=1}^m \bar{r}(X_{i_{\ell}}) \mid X_{i_1}, \ldots, X_{i_v} \right]^2 \right].\] (H8)

Next, since \((i_1, \ldots, i_m)\) are distinct, the variables \(X_{i_1}, \ldots, X_{i_m}\) are independent. Hence we have that

\[\mathbb{E}_Q \left[ \prod_{\ell=1}^m \bar{r}(X_{i_{\ell}}) \mid X_{i_1}, \ldots, X_{i_v} \right] = \left( \prod_{\ell=1}^v \bar{r}(X_{i_{\ell}}) \right) \prod_{\ell=v+1}^m \mathbb{E}_Q[\bar{r}(X_{i_{\ell}})]\] (H9)

\[= \left( \prod_{\ell=1}^v \bar{r}(X_{i_{\ell}}) \right),\]
where the last equality follows because:

\[ \mathbb{E}_Q[\bar{r}(X_{i})] = \int \bar{r}(x)q(x)\,d\mu(x) = \int p(x)\,d\mu(x) = 1. \]

Next, combining (H9) and (H8) we get that

\[ \zeta_v \leq \mathbb{E}_Q \left[ \left( \prod_{\ell=1}^{v} \bar{r}(X_{i_{\ell}}) \right)^2 \right] 
= \prod_{\ell=1}^{v} \mathbb{E}_Q[\bar{r}(X_{i_{\ell}})^2] 
= \mathbb{E}_Q[\bar{r}(X_{i_{1}})^2]^v. \]

Here, we again use the independence of the distinct terms. Plugging this into (H7), we get

\[ \forall \omega \left( \frac{1}{n} \sum_{i_1 \ldots i_{n-m} \text{ distinct}} \left( \prod_{\ell=1}^{m} \bar{r}(X_{i_{\ell}}) \right) \delta_m \right) 
\leq mm^{-1} \sum_{\ell=1}^{m} m\ell n - mm - \ell \mathbb{E}_Q[\bar{r}(X_{i_{1}})^2]. \]

By (A2), \( \mathbb{E}_Q[\bar{r}(X_{i_{1}})^2] < \infty \), so Lemma 2 implies that this converges to 0 for \( n \to \infty \). This shows that the variance converges to zero in both cases (H5) and (H4), which completes the proof of Lemma 1. \( \square \)

**Lemma 2** Let \( m = o(\sqrt{n}) \) as \( n \) goes to infinity. Then for any \( K \geq 0 \), it holds that

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{\ell=1}^{m} \binom{m}{\ell} \binom{n-m}{m-\ell} K^\ell = 0. \] \( \text{(H10)} \)

**Remark 2** The Chu–Vandermonde identity states that

\[ \sum_{\ell=0}^{m} \binom{m}{\ell} \binom{n-m}{m-\ell} = 1. \]

In the light of this identity, one may be surprised that when including the exponentially growing term, \( K^\ell \), the sum vanishes. The reason is that the summation in (H10) starts at \( \ell = 1 \), not \( \ell = 0 \), and since \( n \) grows at least quadratically in \( m \), \( \binom{n-m}{m-\ell} \) for \( \ell = 0 \) dominates all the other summands as \( n \) (and thereby also \( m \)) approaches \( \infty \).

**Proof.** Denote by \( s_{\ell} \) the \( \ell \)th summand, i.e.,

\[ s_{\ell} := \binom{m}{\ell} \binom{n-m}{m-\ell} K^\ell. \]

Denote by \( s_{\ell} \) the \( \ell \)th summand, i.e.,

\[ s_{\ell} := \binom{m}{\ell} \binom{n-m}{m-\ell} K^\ell. \]
It then holds for all $\ell \in \{1, \ldots, m-1\}$ that

$$s_{\ell+1} = \frac{\binom{m}{\ell+1}(n-m)K_{\ell+1}}{\binom{m}{\ell}(n-m)K^\ell} \leq \frac{m!}{(\ell + 1)! (m - \ell - 1)! (n - 2m + \ell + 1)!} \frac{(n-m)!}{\ell!(m-\ell)!(n-2m+\ell)!} K = \frac{(m-\ell)^2}{(\ell+1)(n-2m+\ell+1)} K \leq \frac{m^2}{2(n-2m+2)} K.$$ 

Since, by assumption, $m = o(\sqrt{n})$, this converges to 0 as $n$ goes to infinity. In particular, there exists a constant $c \in (0, 1)$ such that for $n$ sufficiently large it holds for all $\ell \in \{1, \ldots, m-1\}$ that $s_\ell \leq s_1 c^{\ell-1}$. This implies that $s_\ell \leq s_1 c^{\ell-1}$, and hence also

$$\sum_{\ell=1}^m s_\ell \leq s_1 \sum_{\ell=1}^m c^{\ell-1} \leq s_1 \frac{1}{1-c},$$

where for the last inequality we used the explicit solution of a geometric sum. We now conclude the proof by explicitly bounding (H10) as follows:

$$\frac{1}{n} \sum_{\ell=1}^m \binom{m}{\ell} \frac{(n-m)}{(m-\ell)} K^\ell = \frac{1}{n} \sum_{\ell=1}^m s_\ell \leq \frac{1}{1-c} \frac{s_1}{n} \frac{m}{n} \binom{n-m}{m-1} = \frac{K m^2}{1-c} \frac{n}{n-1} \binom{n-m}{m-1} \binom{n}{m},$$

where in the last equation we use the relation $\binom{n}{m} = \frac{n}{m} \binom{n-1}{m-1}$. Using that by assumption $\lim_{n \to \infty} \frac{m^2}{n} = 0$ and that $\frac{n-m}{n-1} \leq 1$, it immediately follows that (H11) converges to zero. This completes the proof of Lemma 2. □
H.2 Proof of Theorem 2

Proof. We explicitly construct an example hypothesis test for which the worst case rate is achieved. We construct a target and observation density on \([0, \infty)\). First, for fixed \(\alpha \in (0, 1)\) and all \(v \in \mathbb{N} \setminus \{0\}\) define

\[ c_v := (1 - \alpha)^{\frac{1}{v}} \quad \text{and} \quad p_v := 1 - (v + 1)^{-\varepsilon}, \]

and \(c_0 := 0\) and \(p_0 := 0\), with \(\varepsilon \in (0, \frac{\ell}{e - 1})\) to be chosen below. Then, for all \(v \in \mathbb{N}\) define

\[ f_v := c_{v+1} - c_v \quad \text{and} \quad g_v := p_{v+1} - p_v. \]

Using these sequences, we define the following two densities:

1. Target density (cdf is denoted by \(F\)): For all \(x \in \mathbb{R}\), we define

\[ f(x) := \sum_{i=0}^{\infty} \mathbb{1}_{\{v \leq x < v+1\}} f_v, \]

2. Observation density (cdf is denoted by \(G\)): For all \(x \in \mathbb{R}\), we define

\[ g(x) := \sum_{i=0}^{\infty} \mathbb{1}_{\{v \leq x < v+1\}} g_v. \]

As \(\lim_{v \to \infty} c_v = \lim_{v \to \infty} p_v = 1\), these functions are indeed densities. Moreover, one can verify that for all \(m \in \mathbb{N}\), we have

\[ F(m) = c_m \quad \text{and} \quad G(m) = p_m. \]

A visualisation of the densities is given in Figure H.1.

Finally, if we define, for all \(x \geq 0\), \(r(x) := \frac{f(x)}{g(x)}\) then we get

\[ \mathbb{E}_x [r(X)^{\ell}] = \int_0^{\infty} r(x)^{\ell} g(x) \, dx = \sum_{v=1}^{\infty} \left( \frac{f_v}{g_v} \right)^{\ell} g_v = \sum_{v=1}^{\infty} \frac{(c_{v+1} - c_v)^{\ell}}{(p_{v+1} - p_v)^{\ell - 1}}. \]

This series converges for all possible parameter choices \(\varepsilon \in (0, \frac{\ell}{e - 1})\) because

(a) \((c_{v+1} - c_v) \sim -\log (1 - \alpha)v^{-2}\) as \(v \to \infty\) and

(b) \((p_{v+1} - p_v) \sim \varepsilon v^{-\varepsilon + 1}\) as \(v \to \infty\).

Indeed, both results follow from the mean value theorem as follows: First, for (a) applying the mean value theorem to \(x \mapsto (1 - \alpha)^{1/x}\) implies that for all \(v \in \mathbb{N}\) there exists \(\xi_v \in [v, v + 1]\) such that

\[ \frac{c_{v+1} - c_v}{v + 1 - v} = -\log (1 - \alpha) \frac{(1 - \alpha)^{1/\xi_v}}{\xi_v^2}. \]
We therefore get
\[
\lim_{v \to \infty} (c_{v+1} - c_v)v^2 = -\log(1 - \alpha) \lim_{v \to \infty} \frac{v^2(1 - \alpha)^{1/\xi_v}}{v^2} = -\log(1 - \alpha).
\]

Similarly, for (b), we apply the mean value theorem to \( x \mapsto 1 - (x + 1)^{-\varepsilon} \) to get that for all \( v \in \mathbb{N} \) there is a \( \xi_v \in [v, v + 1] \) such that
\[
\frac{p_{v+1} - p_v}{v + 1 - v} = \varepsilon (\xi_v + 1)^{-v+1}.
\]

Again taking the limits we get
\[
\lim_{v \to \infty} (p_{v+1} - p_v)v^{v+1} = \varepsilon \lim_{v \to \infty} (\xi_v + 1)^{v+1}v^{v+1} = \varepsilon.
\]

This completes the proofs of (a) and (b).

Now, consider the null hypothesis
\( H_0 := \{P_f\} \).

For all \( m \in \mathbb{N} \), we define the test \( \phi_m : [0, \infty)^m \to \{0, 1\} \) for all \( x_1, \ldots, x_n \in [0, \infty) \) by
\[
\phi_m(x_1, \ldots, x_m) := \mathbb{1}_{\{\max(x_1, \ldots, x_m) \leq m\}}.
\]

Then, it holds that
\[
P_f(\phi_m(X_1, \ldots, X_m) = 0) = P_f(\max(X_1, \ldots, X_m) > m)
= 1 - P_f(X_i \leq m)^n
= 1 - F(m)^n
= 1 - c_m^n
= 1 - [(1 - \alpha)^{1/\xi}]^m
= \alpha.
\]

Hence, \( \phi_m \) achieves valid level in the target distribution \( f \). Our goal is now to show that any resampling procedure for testing under distributional shifts with \( m = n^q \) and \( q > \frac{\varepsilon - 1}{\ell - 1} \) cannot achieve asymptotic level. Let \( \Psi^m \) be the resampling scheme from the theorem that outputs a (not necessarily distinct) sample of size \( m = n^q \).

Then, it holds that
\[
P_g(\phi_m(\Psi^m(X_1, \ldots, X_n)) = 1) = P_g(\max(\Psi^m(X_1, \ldots, X_n)) \leq m)
\geq P_g(\max(X_1, \ldots, X_n) \leq m)
= P_g(X_i \leq m)^n
= G(m)^n
= p_m^n
= (1 - (m + 1)^{-\varepsilon})^n
= \exp(n \log(1 - (m + 1)^{-\varepsilon})).
\]

Taylor expanding \( x \mapsto \log(x) \) in \( x_0 = 1 \) yields
\[
\log(x) = \log(x_0) + \frac{1}{x_0}(x - x_0) + \frac{1 - 1}{2} \frac{1}{x_0^2} (x - x_0)^2
\]
for an $\xi_x \in (x, 1)$. Plugging in $x = 1 - (m + 1)^{-\varepsilon}$, we get
\[
\log(1 - (m + 1)^{-\varepsilon}) = -(m + 1)^{-\varepsilon} = \frac{1}{2\xi_m^2}(m + 1)^{-2\varepsilon},
\]
where $\xi_x$ is lower bounded by $(1 - (m + 1)^{-\varepsilon})$, and so $\xi_m \to 1$ and thus $\xi_m^2 \to 1$ for $m \to \infty$. Next, observe that
\[
n(m + 1)^{-\varepsilon} \leq n m^{-\varepsilon} = n^{1-q \varepsilon}.
\]
Now if we select $\frac{q-1}{q} > \varepsilon > 1/q$ (this is always possible because $q > 1/2$), we have that $q \varepsilon > 1$, and it holds that $\lim_{n \to \infty} n^{1-q \varepsilon} = 0$. For the same reason, we have $n(m + 1)^{-2\varepsilon} \leq n m^{-\varepsilon} \to 0$ as $n \to \infty$. Combining this with the (H12), we get
\[
\lim_{n \to \infty} n \log(1 - (m + 1)^{-\varepsilon}) = 0
\]
and thus
\[
\lim_{n \to \infty} \mathbb{P}_g(\phi_m(\Psi^m(X_1, \ldots, X_n)) = 1) \geq 1.
\]
This completes the proof of Theorem 2. $\square$

H.3 Proof of Theorem 3

Proof. The proof is similar to the proof of Theorem 1 but we will need to adjust for the estimation of the distributional shift factor. In particular, we will reprove the results in Lemma 1 when using the estimator $\hat{r}_n$. 

Figure H.1. Visualisation of densities in the proof of Theorem 2 with $\varepsilon = 1.9$. The tail of the target density eventually becomes larger than that of the observation density.
Fix any $P \in H_0$ and let $Q \in \tau^{-1}(\{P\})$. Denote by $p$ and $q$ their respective densities with respect to the dominating measure $\mu$. We begin by recalling the details for the sample splitting procedure described in Algorithm 3: $X_n$ is split into two disjoint data sets $X_{n_1}$ and $X_{n_2}$ of sizes $n_1, n_2$, where $n_1 + n_2 = n$. The assumptions $n_1 = \sqrt{n_2}$ and $m = o(\min(n', \sqrt{n})$), ensure that $m = o(n_1^a)$ and $m = o(\sqrt{n_2})$. We use $X_{n_1}$ to fit an estimator $\hat{r}_n$ of $r_q$ and then use $X_{n_2}$ for the resampling. When taking expectations over $X_{n_1}$, we use the notation $E_{Q_1}$. Similarly, $E_{Q_2}$ denotes an expectation over $X_{n_2}$. We write $E_Q$ when taking expectations with respect to the entire sample $X_n$. Let $I_1 := \{1, \ldots, n_1\}$ and $I_2 := \{n_1 + 1, \ldots, n_1 + n_2\}$ be the indices of $X_{n_1}$ and $X_{n_2}$, respectively.

Using the same argument as we used to derive (H1) in the proof of Theorem 1, we get that

\[
\mathbb{P}_Q(\varphi_m(\Psi^{n_2}(X_{n_2}, U) = 1)) = E_Q \left[ \frac{1}{n_2!} \sum_{(i_1, \ldots, i_m) \text{ distance from } I_2} \left( \prod_{i=1}^m \hat{r}_n(X_{i_1}) \right) \mathbb{I}_{\{\varphi_m(X_{i_1}, \ldots, X_{i_m}) = 1\}} \right],
\]

(H13)

where $X_{i_1}$ are observations from $X_{n_2}$. As in the proof of Lemma 1, we prove the convergence in probability of the numerator and denominator in (H13) separately. Again, we do this in two steps: (A) We show that the means converge to the desired quantity and (B) we show that the variances converge to zero. First, we show the following intermediate result.

**Intermediate result:** Let $\varepsilon(n_1) := \sup_{x \in X} E_{Q_1} \left| \left( \frac{\hat{r}_n(x)}{r_q(x)} \right)^{n_1} - 1 \right|$ and consider a sequence $i_1, \ldots, i_m$ from the indices of $X_2$. Then, for $n_1$ sufficiently large and using Jensen’s inequality, it holds $Q_{n_2}$-a.s. that

\[
\left| E_{Q_1} \left[ \prod_{i=1}^m \hat{r}_n(X_{i_1}) \right] - 1 \right| \leq E_{Q_1} \left[ \prod_{i=1}^m \frac{\hat{r}_n(X_{i_1})}{r_q(X_{i_1})} \right] - 1 \]

\[
\leq \sup_{x_1, \ldots, x_m \in X} E_{Q_1} \left[ \prod_{i=1}^m \frac{\hat{r}_n(x_{i_1})}{r_q(x_{i_1})} - 1 \right]
\]

\[
\leq \sup_{x \in X} E_{Q_1} \left[ \left( \frac{\hat{r}_n(x)}{r_q(x)} \right)^m - 1 \right]
\]

\[
\leq \sup_{x \in X} E_{Q_1} \left[ \left( \frac{\hat{r}_n(x)}{r_q(x)} \right)^{n_1} - 1 \right]
\]

\[
= \varepsilon(n_1).
\]

The last inequality holds because by assumption, $m = o(n_1^a)$ when $n_1 \to \infty$, so for $n_1$ sufficiently large, $n_1^a > m$. For any $m, k \in \mathbb{N}$ we have $c^{m+k} \geq c^m \geq 1$ if $c > 1$ and $c^{m+k} \leq c^m \leq 1$ if $0 \leq c \leq 1$, and in either case it holds that $|c^{m+k} - 1| \geq |c^m - 1|$. Similarly, for $i_1, \ldots, i_m$ and $\hat{r}_1, \ldots, \hat{r}_m$ from $I_2$, we get $Q_{n_2}$-a.s. that

---

13 When $n_1 = \sqrt{n_2}$ and $n_1 + n_2 = n$, we have $n = n_1^{2+a} + n_1$ and $n = n_2 + n_2^{1/(2a)}$. If $a > 1/2$, we have $m = o(\sqrt{n}) = o(\sqrt{n_2}) = o(n_1^a)$, and $m = o(\sqrt{n}) = o(\sqrt{n_2})$. Similar arguments apply if $a < 1/2$. 
Furthermore, using the intermediate result (H14), we get the following upper bound:

\[
\begin{align*}
\mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_{\ell}})\right) \left(\prod_{\ell=1}^{m} r_q(X_{i_{\ell}})\right) - 1\right] &\leq \mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_{\ell}})\right) \left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_{\ell}})\right) - 1\right] \\
&\leq \sup_{x \in X} \mathbb{E}_{Q_1}\left[\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{2m} - 1\right] \\
&\leq \sup_{x \in X} \mathbb{E}_{Q}\left[\left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{n_1^q} - 1\right] \\
&= \varepsilon(n_1),
\end{align*}
\]

using that for \( n_1 \) sufficiently large, \( n_1^q > 2m \). This concludes the intermediate result.

Before showing parts A and B, we introduce some notation. Depending on whether we consider the numerator or the denominator case, we define either \( \delta_m := \varepsilon_{\{\varepsilon_m(X_{i_1}, \ldots, X_{i_m}) = 1\}} \) or \( \delta_m := 1 \). Furthermore, we introduce for any function \( r: \mathcal{X} \to (0, \infty) \) the following random variable:

\[
M(r) := \frac{1}{n_2!} \sum_{\{i_1, \ldots, i_m\} \text{ distinct from } i_2} \left( \prod_{\ell=1}^{m} r(X_{i_{\ell}}) \right) \delta_m.
\]

**Part A (means):** Since \( \hat{r}_{n_1}(x) = r_q(x) \frac{\hat{r}_m(x)}{r_q(x)} \), using the independence between \( X_{n_1} \) and \( X_{n_2} \) we get that

\[
\begin{align*}
\mathbb{E}_{Q}[M(\hat{r}_{n_1})] &= \frac{1}{n_2!} \sum_{\{i_1, \ldots, i_m\} \text{ distinct from } i_2} \mathbb{E}_{Q_I}\left[\left(\prod_{\ell=1}^{m} r_q(X_{i_{\ell}})\right) \left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_{\ell}})\right) \delta_m\right] \\
&= \frac{1}{n_2!} \sum_{\{i_1, \ldots, i_m\} \text{ distinct from } i_2} \mathbb{E}_{Q_I}\left[\left(\prod_{\ell=1}^{m} r_q(X_{i_{\ell}})\right) \mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^{m} \hat{r}_{n_1}(X_{i_{\ell}})\right) \delta_m\right]\right].
\end{align*}
\]

We emphasise that the expectation \( \mathbb{E}_{Q_I} \) only averages over the randomness in estimating \( \hat{r} \), and does take expectations over \( X_{i_2} \) which is drawn from \( Q_{n_2} \).

Furthermore, using the intermediate result (H14), we get the following upper bound:

\[
\mathbb{E}_{Q}[M(\hat{r}_{n_1})] \leq \mathbb{E}_{Q_1}(M(r_q))(1 + \varepsilon(n_1))
\]

and lower bound

\[
\mathbb{E}_{Q}[M(\hat{r}_{n_1})] \geq \mathbb{E}_{Q_1}(M(r_q))(1 - \varepsilon(n_1)).
\]

Since \( m = o(\sqrt{n_2}) \), we can apply Lemma 1 (a) and (b) to get that the means \( \mathbb{E}_{Q}[M(\hat{r}_{n_1})] \) of the denominator and numerator converge to the desired values.
H.4 Proof of Theorem 4

**Proof of Theorem 4.** The first part of the proof follows that of Theorem 1. Let $p$ and $q$ denote the respective densities of $P$ and $Q$ with respect to the dominating measure $\mu$. Recall that we call a sequence $(i_1, \ldots, i_m)$ distinct if for all $\ell \neq \ell'$ we have $i_\ell \neq i_{\ell'}$. The resampling scheme $\Psi^{\text{DRPL}}$ samples from the space of distinct sequences $(i_1, \ldots, i_m)$, where every sequence has probability $\nu_{\Psi^{\text{DRPL}}}(i_1, \ldots, i_m) \propto \prod_{\ell=1}^m r(X_{i_\ell})$. The normalisation constant is the sum over the weights in the entire space of distinct sequences, that is,

$$\nu_{\Psi^{\text{DRPL}}}(i_1, \ldots, i_m) = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{\{i_1, \ldots, i_m\} \text{ distinct}} \prod_{\ell=1}^m r(X_{i_\ell})}.$$

Thus, taking an expectation involving $\varphi_m(\Psi^{\text{DRPL}}_{\text{DRPL}}(X, U))$, amounts to evaluating $\varphi_m$ in all distinct sequences $X_{i_1}, \ldots, X_{i_m}$ and weighting with the probabilities $\nu_{\Psi^{\text{DRPL}}}(i_1, \ldots, i_m)$.

$$\mathbb{P}_Q(\varphi_m(\Psi^{\text{DRPL}}_{\text{DRPL}}(X, U)) = 1)$$

$$= \mathbb{E}_Q \left[ \frac{1}{n!} \sum_{\{i_1, \ldots, i_m\} \text{ distinct}} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \delta_{\nu_{\Psi^{\text{DRPL}}}(i_1, \ldots, i_m)=1} \right], \quad (H16)$$

Part B (variances): Next, we show that both for the numerator and denominator the variance converges to zero. To this end, we expand the second moment as follows:

$$\mathbb{E}_Q\left[ (\hat{\varphi}_n)^2 \right] = \mathbb{E}_Q\left[ \frac{1}{n_2!} \prod_{(i_1, \ldots, i_m) \text{ distinct from } 1} \prod_{i_i' \neq i_i} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \delta_m \delta_m' \right]$$

$$= \mathbb{E}_Q \left[ \frac{1}{n_2!} \prod_{(i_1, \ldots, i_m) \text{ distinct from } 1} \prod_{i_i' \neq i_i} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \delta_m \delta_m' \right]$$

$$\mathbb{E}_Q \left[ \prod_{i_i' \neq i_i} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \delta_m \delta_m' \right].$$

Here $\delta_m := \{\nu_m(X_{i_1}, \ldots, X_{i_m})=1\}$. Using the intermediate result (H14), we get the following upper bound:

$$\mathbb{E}_Q\left[ (\hat{\varphi}_n)^2 \right] \leq \mathbb{E}_Q\left[ (\varphi_n)^2 \right] (1 + \epsilon(n_1))$$

and lower bound

$$\mathbb{E}_Q\left[ (\hat{\varphi}_n)^2 \right] \geq \mathbb{E}_Q\left[ (\varphi_n)^2 \right] (1 - \epsilon(n_1)).$$

In Lemma 1 (c) and (d), we have shown that $\lim_{n \to \infty} \mathbb{E}_Q(\varphi_n) = 0$. Hence, combining these bounds on the second moment with the above bounds on the first moment shows that also $\lim_{n \to \infty} \mathbb{E}_Q(\hat{\varphi}_n) = 0$. This completes the proof of Theorem 3. □
where we divide by the number of distinct sequences \( \frac{n!}{(n-m)!} \) in both numerator and denominator.

Let \( c(n, m) \) and \( d(n, m) \) be the numerator and denominator terms of (H16), i.e.,

\[
c(n, m) := \frac{1}{n!} \sum_{(i_1, \ldots, i_m) \text{ distinct}} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \mathbb{1}_{\{\phi_m(X_{i_1}, \ldots, X_{i_m}) = 1\}},
\]

\[
d(n, m) := \frac{1}{(n-m)!} \sum_{(j_1, \ldots, j_m) \text{ distinct}} \prod_{\ell=1}^m r(X_{j_\ell}).
\]

Define for all \( \delta > 0 \) the set \( A_\delta := \{ d(n, m) \geq 1 - \delta \} \). It holds for all \( \delta \in (0, 1) \) that

\[
E_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = E_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{A_\delta} \right] + E_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{A_\delta^c} \right]
\]

\[
\leq E_Q \left[ \frac{c(n, m)}{1 - \delta} \mathbb{1}_{A_\delta} \right] + E_Q \left[ \frac{c(n, m)}{d(n, m)} \mathbb{1}_{A_\delta^c} \right]
\]

\[
\leq E_Q \left[ \frac{c(n, m)}{1 - \delta} \right] + P_Q(A_\delta^c)
\]

\[
= \frac{1}{1 - \delta} P_Q(\phi_m(X_1, \ldots, X_m) = 1) + P_Q(A_\delta^c),
\]

where we used that \( \frac{c(n, m)}{d(n, m)} \leq 1 \) and Lemma 1 (a). Further, by applying Cantelli’s inequality to \( P_Q(A_\delta^c) \), it follows that

\[
P_Q(A_\delta^c) \leq \frac{\text{VAR}_Q(d(n, m))}{\text{VAR}_Q(d(n, m)) + \delta^2}.
\]

Finally, we can apply (H7),

\[
V(n, m) := \text{VAR}_Q(d(n, m))
\]

\[
= \text{VAR} \left( \frac{1}{n!} \sum_{(i_1, \ldots, i_m) \text{ distinct}} \left( \prod_{\ell=1}^m r(X_{i_\ell}) \right) \right)
\]

\[
= \left( \frac{n}{m} \right)^{-1} \sum_{\ell=1}^m \binom{m}{\ell} \binom{n-m}{m-\ell} \left( E_Q \left[ r(X_{i_\ell})^2 \right] - 1 \right),
\]

where we use that \( \zeta_v \) [used in (H7)] is given by

\[
\zeta_v = \mathbb{V}_Q \left( \mathbb{E}_Q \left[ \prod_{\ell=1}^m r(X_{i_\ell}) \mid X_{i_1}, \ldots, X_{i_m} \right] \right)
\]

\[
= \mathbb{V}_Q \left( \left( \prod_{\ell=1}^v r(X_{i_\ell}) \right) \mathbb{E}_Q \left[ \prod_{\ell=v+1}^m r(X_{i_\ell}) \right] \right)
\]

\[
= \mathbb{V}_Q \left( \prod_{\ell=1}^v r(X_{i_\ell}) \right)
\]

\[
= \mathbb{E}_Q[r(X_{i_1})^2] - 1.
\]
Plugging in this upper bound for $P_Q(A_\delta)$ yields

$$P_Q(\varphi_m(\Psi_{\text{DPL}}(X_n, U)) = 1) = \frac{1}{1 - \delta} P_p(\varphi_m(X_1, \ldots, X_m) = 1)
+ \frac{V(n, m)}{V(n, m) + \delta^2}.$$

Since $\delta \in (0, 1)$ was arbitrary, the theorem statement follows. \hfill \Box

H.5 Proof of Theorem 5

Proof. We adjust part of the proof of Theorem 1 to the uniform case. Again, let $c(n, m)$ and $d(n, m)$ be the numerator and denominator terms of (H1), i.e.,

$$c(n, m) := \frac{1}{n!} \sum_{\substack{|\{\ell, \ldots, m\}| = n \text{ distinct}}} \frac{1}{m!} \prod_{\ell=1}^{m} r(X_{i\ell}),$$

$$d(n, m) := \frac{1}{n!} \sum_{\substack{|\{\ell, \ldots, m\}| = n \text{ distinct}}} m! \prod_{\ell=1}^{m} r(X_{i\ell}).$$

We want to show that $\limsup_{n \to \infty} \sup_{Q \in \mathcal{E}(H_0)} E_Q \left[ \frac{c(n, m)}{d(n, m)} \right] \leq \alpha \varphi$. To see this, define for all $\delta > 0$ the set $A_\delta := \{|d(n, m) - 1| \leq \delta\}$, and take any $P \in H_0$ and $Q \in r^{-1}(\{P\})$. It holds for all $\delta \in (0, 1)$ that

$$E_Q \left[ \frac{c(n, m)}{d(n, m)} \right] = E_Q \left[ \frac{c(n, m)}{d(n, m)} 1_{A_\delta} \right] + E_Q \left[ \frac{c(n, m)}{d(n, m)} 1_{A_\delta^c} \right]$$

$$\leq E_Q \left[ \frac{c(n, m)}{d(n, m)} 1_{A_\delta} \right] + E_Q \left[ \frac{c(n, m)}{d(n, m)} 1_{A_\delta^c} \right]$$

$$\leq E_Q \left[ \frac{c(n, m)}{1 - \delta} \right] + \frac{P_Q(A_\delta^c)}{1 - \delta}$$

$$= \frac{1}{1 - \delta} P_p(\varphi_m(X_1, \ldots, X_m) = 1) + P_Q(A_\delta^c),$$

where we used that $\frac{c(n, m)}{d(n, m)} \leq 1$ and that $E_Q[c(n, m)] = P_p(\varphi_m(X_1, \ldots, X_m) = 1)$, as shown in Lemma 1 (a). Further, given the uniform bound on the weights, combining Chebyshev’s inequality with Lemma 1 (b) and (d) leads to $\lim_{n \to \infty} \sup_{Q \in \mathcal{E}(H_0)} P_Q(A_\delta^c) = 0$. Hence, using that $\varphi$ has uniform asymptotic level $\alpha \varphi$ we have shown for all $\delta \in (0, 1)$ that

$$\limsup_{n \to \infty} \sup_{Q \in \mathcal{E}(H_0)} E_Q \left[ \frac{c(n, m)}{d(n, m)} \right] \leq \frac{1}{1 - \delta} \alpha \varphi.$$

Using that $\delta \in (0, 1)$ is arbitrary, completes the proof of Theorem 5. \hfill \Box

H.6 Proof of Corollary 3

Proof. We have

$$P_Q(\varphi_m(\Psi_{\text{REPL}}(X_n, U)) = 1)$$

$$= P_Q(\varphi_m(\Psi_{\text{REPL}}(X_n, U)) = 1 | \Psi_{\text{REPL}}(X_n, U) \text{ distinct}) P_Q(\Psi_{\text{REPL}}(X_n, U) \text{ distinct})$$

$$+ P_Q(\varphi_m(\Psi_{\text{REPL}}(X_n, U)) = 1 | \Psi_{\text{REPL}}(X_n, U) \text{ not distinct}) P_Q(\Psi_{\text{REPL}}(X_n, U) \text{ not distinct}).$$
This converges to the same limit as \( P_Q(\Psi_{\text{REPL}}^m(X_n, U)) = 1 \) \( P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ distinct}) \) (because \( P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ distinct}) \) converges to 1, see Proposition 2), which, as we argue in Appendix D, equals \( P_Q(\psi_m(\Psi_{\text{DRPL}}^m(X_n, U)) = 1 \). The result then follows from Theorem 1. □

H.7 Proof of Proposition 2

Proof. When sampling with replacement, \( \Psi_{\text{REPL}}^m(X_n, U) \) contains nondistinct draws with positive probability (assuming, wlog, that \( m \) is not 1). Yet, we show that \( P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ distinct}) \) approaches 1 as \( m \to \infty \). By assumption \( p = r(q) \), so \( p(x) \propto r(x) q(x) \). Let \( \bar{r} \) be the normalised version of \( r \) satisfying, for all \( x \), \( p(x) = \bar{r}(x) q(x) \). The probability \( \psi_{(i_1, \ldots, i_m)} \) of drawing a sequence \( X_{i_1}, \ldots, X_{i_m} \) is defined by (10) as the product of weights \( r \):

\[
\psi_{(i_1, \ldots, i_m)} = \frac{\prod_{k=1}^{m} r(X_{i_k})}{\sum_{(j_1, \ldots, j_m)} \prod_{k=1}^{m} r(X_{j_k})},
\]

where the sum over \( (j_1, \ldots, j_m) \) in the denominator is over all sequences of length \( m \) (including distinct and nondistinct sequences). The probability of drawing a nondistinct sequence equals the sum of the weights corresponding to all nondistinct sequences \( \psi_{(i_1, \ldots, i_m)} \). Therefore,

\[
P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ not distinct})
= \mathbb{E}_Q \left[ \sum_{(i_1, \ldots, i_m) \text{ not distinct}} \psi_{(i_1, \ldots, i_m)} \right]
= \mathbb{E}_Q \left[ \sum_{(i_1, \ldots, i_m) \text{ not distinct}} \frac{\prod_{k=1}^{m} \bar{r}(X_{i_k})}{\sum_{(j_1, \ldots, j_m)} \prod_{k=1}^{m} \bar{r}(X_{j_k})} \right] \tag{H17}
= \mathbb{E}_Q \left[ \frac{\prod_{k=1}^{m} \bar{r}(X_{i_k})}{\sum_{(j_1, \ldots, j_m) \text{ not distinct}} \prod_{k=1}^{m} \bar{r}(X_{j_k})} \right]
= \mathbb{E}_Q \left[ \frac{1}{m^m} \sum_{(i_1, \ldots, i_m) \text{ not distinct}} \prod_{k=1}^{m} \bar{r}(X_{i_k}) + \frac{1}{m^m} \sum_{(i_1, \ldots, i_m) \text{ distinct}} \prod_{k=1}^{m} \bar{r}(X_{i_k}) \right].
\]

Observe that this expectation is taken both over \( X_n \) and \( U \). By Lemma 3, the numerator of (H17) (which equals the first term in the denominator) converges to 0 in \( L^1 \). The second term in the denominator converges in probability to 1 by Lemma 1 [this requires (A2), which is implied by (A3)]; thus, the entire denominator converges to 1 in probability. By Slutsky’s lemma, the entire fraction (inside the mean) converges to 0 in probability. Since the fraction is lower bounded by 0 and upper bounded by 1, convergence in probability implies convergence of the mean (see the proof of Theorem 1 for an argument for this), and it follows that \( P_Q(\Psi_{\text{REPL}}^m(X_n, U) \text{ not distinct}) \to 0 \). □

Lemma 3 (Nondistinct draws). Let \( P \in \mathcal{P} \) and \( Q \in \mathcal{Q} \) be distributions with densities \( p \) and \( q \) with respect to a dominating measure \( \mu \). Let \( \bar{r}: \mathcal{X} \to (0, \infty) \) satisfy for
all $x \in \mathcal{X}$ that $p(x) = \tilde{q}(x)q(x)$. Then, under (A1) and (A3) it holds that

$$ \lim_{n \to \infty} \mathbb{E}_Q \left[ \frac{1}{nm} \sum_{(i_1, \ldots, i_m) \text{ not distinct}} m \prod_{\ell=1}^m \tilde{r}(X_{i_{\ell}}) \right] = 0. $$

In particular, since the integrand is nonnegative, this implies that

$$ \frac{1}{n^m} \sum_{(i_1, \ldots, i_m) \text{ not distinct}} m \prod_{\ell=1}^m \tilde{r}(X_{i_{\ell}}) \overset{L^1}{\to} 0 \quad \text{as} \ n \to \infty. $$

**Proof.** We first rewrite the sum using the number $k$ of distinct draws, i.e., we consider cases, in which there are $k$ distinct elements among $i_1, \ldots, i_m$. The number $k$ is at least 1 and, since not all draws are distinct, at most $m - 1$. For fixed $k$, we then further sum over the numbers $r_1, \ldots, r_k$ of occurrences of each index, i.e., $j_{\ell}$ appears $r_{\ell}$ times.

$$ \sum_{(i_1, \ldots, i_m) \text{ not distinct}} m \prod_{\ell=1}^m \tilde{r}(X_{i_{\ell}}) = \sum_{k=1}^{m-1} \sum_{(i_1, \ldots, i_m) \text{ with } k \text{ distinct entries}} \sum_{(i_1, \ldots, i_m) \text{ with } r_1 > 0 \text{ times, } r_1 + \cdots + r_k = m} \prod_{\ell=1}^k \tilde{r}(X_{i_{\ell}})^{r_{\ell}}. $$

Using the independence across distinct observations, this implies that

$$ \mathbb{E}_Q \left[ \sum_{(i_1, \ldots, i_m) \text{ not distinct}} m \prod_{\ell=1}^m \tilde{r}(X_{i_{\ell}}) \right] = \sum_{k=1}^{m-1} \sum_{(i_1, \ldots, i_m) \text{ with } k \text{ distinct entries}} \sum_{(i_1, \ldots, i_m) \text{ with } r_1 > 0 \text{ times, } r_1 + \cdots + r_k = m} \mathbb{E}_Q \left[ \tilde{r}(X_{i_{\ell}})^{r_{\ell}} \right]. \quad (H18) $$

We now use the uniform bound on the weights given in (A3) and the fact that $\mathbb{E}_Q[\tilde{r}(X_i)^t] = \tilde{r}(x_i)q(x_i)q(X_i)^t \mu(x_i) = \int p(x_i)\tilde{r}(x_i)^t \mu(x_i) = \mathbb{E}_P[\tilde{r}(X_i)^{t-1}]$ to get for all $i \in \{1, \ldots, n\}$ and all $t \in \{1, \ldots, m - 1\}$ that

$$ \mathbb{E}_Q[\tilde{r}(X_i)^t] = \mathbb{E}_P[\tilde{r}(X_i)^{t-1}] \leq L^{t-1}. $$

Together with (H18) this results in

$$ \mathbb{E} \left[ \sum_{(i_1, \ldots, i_m) \text{ not distinct}} m \prod_{\ell=1}^m \tilde{r}(X_{i_{\ell}}) \right] \leq m-1 \sum_{k=1}^{m} \sum_{(i_1, \ldots, i_m) \text{ with } k \text{ distinct entries}} \sum_{(i_1, \ldots, i_m) \text{ with } r_1 > 0 \text{ times, } r_1 + \cdots + r_k = m} \mathbb{E}_Q \left[ \tilde{r}(X_{i_{\ell}})^{r_{\ell}} \right] \leq m-1 \sum_{k=1}^{m} \binom{n}{k} \pi(m, k) L^{m-k} \leq m-1 \sum_{k=1}^{m} \binom{n}{k} \tilde{r}(m, k) L^{m-k}, \quad (H19) $$

where $\pi(m, k)$ is the number of words of length $m$ using $k$ letters such that each letter is used at least once.
\[ \tilde{\pi}(m, k) := k^{m-k} \frac{m!}{(m-k)!}. \]

The last inequality holds because we have \( \pi(m, k) \leq \tilde{\pi}(m, k) \): Consider constructing a word of length \( m \) by first distributing one of each of the \( k \) letters (ensuring that each letter is used at least once) among the \( m \) positions, which can be done in \( m!/ (m-k)! \) ways. For the remaining \( m-k \) positions pick any combination of letters, which can be done in \( k^{m-k} \) ways. In total, this two-step procedure has \( \pi(m, k) \) possible outcomes. This enumeration contains all words of length \( m \) using \( k \) letters such that each is used at least once, so \( \pi(m, k) \leq \tilde{\pi}(m, k) \). We do not have equality, because \( \tilde{\pi} \) counts some words several times, but with different intermediate steps. For example if \( k = 2 \) and \( m = 3 \), \( \tilde{\pi} \) counts \((a, _, b) + (_, a, _) \) and \((_, a, b) + (a, _, _) \) as two distinct words, although they both yield \((a, a, b) \); indeed \( \pi(3, 2) = 6 \) and \( \tilde{\pi}(3, 2) = 12 \).

Then, with \( s_k := \binom{n}{k} \tilde{\pi}(m, k) L^{m-k} \) it holds that
\[
\frac{s_{k+1}}{s_k} = \frac{\binom{n}{k+1} \tilde{\pi}(m, k+1) L^{m-k-1}}{\binom{n}{k} \tilde{\pi}(m, k) L^{m-k}} = \frac{1 - \frac{k}{L} - \frac{k+1}{L}}{1 - \frac{k}{L}} \geq \frac{1}{c},
\]
where the inequality follows by using \( k \leq m-1 \) and \( (k+1)/k \geq 1 \). By \((A1) \) (i.e., \( m = o(\sqrt{n}) \)) it holds for \( n \) sufficiently large that \( c > 1 \). Iterating this inequality, we get (again for \( n \) sufficiently large) that \( s_k \leq c^{-(m-1-k)} s_{m-1} \), which we can plug into \((H19) \) to get
\[
\mathbb{E} \left[ \sum_{\ell=1}^{m} \prod_{\ell \neq \text{not distinct}} X_{\ell} \right] \leq \sum_{k=1}^{m-1} s_k \leq s_{m-1} \sum_{k=1}^{m-1} c^{-(m-1-k)} = s_{m-1} \sum_{k=0}^{m-2} c^{-k} \leq s_{m-1} \frac{1}{1-c}. \quad \text{(H20)}
\]
In the last inequality, we use the trivial bound \( \sum_{k=0}^{m-2} c^{-k} < \sum_{k=0}^{\infty} c^{-k} \) and \( 0 < c^{-1} < 1 \). Finally, observe that
\[
n^{-m} s_{m-1} = n^{-m} \binom{n}{m-1} \tilde{\pi}(m, m-1) L.
\]

\[
= n^{-m} \frac{n!}{(m-(m-1))!(m-1)!} \frac{(m-1)^{m-(m-1)} m!}{(m-(m-1))!} L
\]

\[
= n^{-m} \frac{n!}{(n-m)!} \frac{m(m-1)}{(n-m+1)} L,
\]

\[
= \frac{g(n, m)}{\frac{m(m-1)}{(n-m+1)}} L,
\]
which by Lemma 4 converges to zero (by the assumption \( m = o(\sqrt{n}) \)). Therefore, we have that

\[
E_{\eta} \left[ n^{-m} \sum_{\{i_1, ..., i_m\} \text{not distinct}} \prod_{\ell=1}^m \tilde{f}(X_{i_{\ell}}) \right] \leq n^{-m} s_{m-1} \frac{1}{1 - \frac{c}{c}} \rightarrow 0,
\]

which completes the proof of Lemma 3. □

Lemma 4 Define for all \( n, m \in \mathbb{N} \) the function

\[
g(n, m) := \frac{n!}{(n-m)!} n^{-m}.
\]

Then, it holds that

\[
\lim_{n \to \infty} g(n, n^q) = \begin{cases} 0 & \text{if } q \in (\frac{1}{2}, 1) \\ \exp\left(-\frac{1}{2}\right) & \text{if } q = \frac{1}{2} \\ 1 & \text{if } q \in [0, \frac{1}{2}) \end{cases}
\]

Proof. First, apply the Stirling approximation to get for \( n \) sufficiently large that

\[
g(n, m) \sim n^{\frac{n+1}{2}} e^{-n} \cdot (n-m)^{m-n^{-\frac{1}{2}}} e^{-m} \cdot n^{-m}
\]

\[
= n^{n-m+\frac{1}{2}} (n-m)^{m-n^{-\frac{1}{2}}} e^{-m}
\]

\[
= \exp\left\{ \left( n - m + \frac{1}{2} \right) \log(n) + \left( m - n - \frac{1}{2} \right) \log(n - m) - m \right\}.
\]

Next, we look at cases where \( m = n^q \) for some \( q \in [0, 1) \). The above expression can then be simplified further as

\[
g(n, n^q) \sim \exp\left\{ \left( n - n^q + \frac{1}{2} \right) \log(n) + \left( n^q - n - \frac{1}{2} \right) \log(n - n^q) - n^q \right\}
\]

\[
= \exp\left\{ \left( n - n^q + \frac{1}{2} \right) \log(n) + \left( n^q - n - \frac{1}{2} \right) \log(1 - n^{q-1}) - n^q \right\}
\]

\[
= \exp\left\{ \left( n^q - n - \frac{1}{2} \right) \log(1 - n^{q-1}) - n^q \right\}.
\]

Finally, since \( n^{q-1} \rightarrow 0 \) as \( n \) goes to infinity we can use the following Taylor expansion:

\[
\log(1 - n^{q-1}) = -n^{q-1} - \frac{1}{2} n^{2(q-1)} + O(n^{3(q-1)}),
\]

which results in

\[
g(n, n^q) \sim \exp\left\{ \left( n^q - n - \frac{1}{2} \right) \log(1 - n^{q-1}) - n^q \right\}
\]

\[
= \exp\left\{ n^q - n - \frac{1}{2} \left( -n^{q-1} - \frac{1}{2} n^{2(q-1)} + O(n^{3(q-1)}) \right) - n^q \right\}
\]

\[
= \exp\left\{ -n^{2q-1} - \frac{1}{2} n^{3q-2} + n^q + \frac{1}{2} n^{3q-2} + \frac{1}{2} n^{q-1} + \frac{1}{4} n^{2q-2} + O(n^{2q-1}) - n^q \right\}
\]

\[
= \exp\left\{ -\frac{1}{2} n^{2q-1} + O(n^{3q-2}) \right\}.
\]
From this we see that
\[\lim_{n \to \infty} g(n, n^q) = \begin{cases} 
0 & \text{if } q \in (\frac{1}{2}, 1) \\
\exp(-\frac{1}{2}) & \text{if } q = \frac{1}{2} \\
1 & \text{if } q \in [0, \frac{1}{2}).
\end{cases}\]

This completes the proof of Lemma 4. □

H.8 Proof of Corollary 1
As discussed in Section 2.3, the proposed procedure in Section 4.1 can also be used to construct a test for a hypothesis \(H_0^Q\) in the observed domain, satisfying the same theoretical guarantees.

**Corollary 4** (Pointwise level in the observed domain—detailed version). Consider hypotheses \(H_0^Q \subseteq Q\) and \(H_0^P \subseteq P\) in the observational and in the target domain, respectively. Let \(\tau : Q \to P\) be a distributional shift for which there exist a known map \(r : X \to (0, \infty)\) and a set \(A\) satisfying for all \(q \in Q\) and all \(x \in Z\) that \(\tau(q)(x) \propto r(qA)q(x)\), see (5). Assume \(\tau(H_0^Q) \subseteq H_0^P\). Let \(\phi_k\) be a sequence of tests for \(H_0^P\) with pointwise asymptotic level \(\alpha_\phi\). Let \(m = m(n)\) be a resampling size and let \(\psi_n^r\) be the DRPL-based resampling test defined by \(\psi_n^r(X_n, U) = \phi_m(\Psi_{DRPL}(X_n, U))\), see Algorithm 1. Then, if \(m\) satisfies (A1) and all \(Q \in H_0^Q\) satisfy (A2), it holds that
\[\sup_{Q \in H_0^Q} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_{n}^r = 1) \leq \alpha_\phi,\]
i.e., \(\psi_n^r\) satisfies pointwise asymptotic level \(\alpha\) for the hypothesis \(H_0^Q\).

Clearly, the condition \(H_0^Q \subseteq \tau(H_0^P)\) is satisfied when \(H_0^P = \tau(H_0^Q)\). This is the case for the conditional independence test described in Section 3.1, for example.

**Proof.** We have
\[\tau(H_0^Q) \subseteq H_0^P \Rightarrow H_0^Q \subseteq \tau^{-1}(H_0^P)\]
and therefore
\[\sup_{Q \in H_0^Q} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_{n}^r = 1) \leq \sup_{Q \in \tau^{-1}(H_0^P)} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_{n}^r = 1).\]

Since (A2) is satisfied for all \(Q \in H_0^Q\), the statement follows from Theorem 1. This completes the proof of Corollary 1. □

H.9 Proof of Proposition 1

**Proof.** We analyse the output of Algorithm 2. For each \(i \in \{1, \ldots, n\}\), we discard \(X_i\) if \(U_i > r(X_i)/M\), where \(U_i\) is uniform on \((0, 1)\). The probability of the event \(E_i\) that \(X_i\) is not discarded equals
\[q(E_i) = \int q(E_i|x_i)q(x_i)dx_i = \int \frac{r(x_i)}{M}q(x_i)dx_i = \int \frac{c}{M}p(x_i)q(x_i)dx_i = \frac{c}{M},\]
where \(c\) is a constant such that \(r(x)q(x) = cp(x)\), and the conditional density of \(X_i\) given that the sample is not discarded, \(q(x_i|E_i)\), is given by
\[ q(x_i|E_i) = \frac{q(x_i)}{q(E_i)} \quad q(E_i|x_i) = q(x_i) \]

If \(X_{i1}, \ldots, X_{im}\) are the points that are not discarded, this means that \((X_{i1}, \ldots, X_{im})\) is distributed as if it was \(m\) i.i.d. draws from \(p^*\) (where \(m\) is random). In particular, by the assumption that the probability that for all \(k \in \mathbb{Z}: P_P (\varphi_k (Z_k) = 1) = \alpha_p\) for \(k\) i.i.d. samples \(Z_k\) from \(P_P\), it follows that \(P_Q (\varphi_k (X_k, U) = 1) = P_Q (\varphi (X_{i1}, \ldots, X_{im})) = \alpha_p\). \(\square\)

**Appendix I: Analysing (A2) in a linear Gaussian model**

In this section, we show conditions for assumption (A2) to be satisfied when we consider the shift that changes a Gaussian conditional into a marginal, independent Gaussian target distribution.

**Proposition 3**  
Consider a linear Gaussian setting where \(Y = X + \epsilon\) with \(\epsilon \sim N(0, \sigma^2_\epsilon)\) and \(X \sim N(0, \sigma^2_X)\), with \(\sigma_\epsilon, \sigma_X\) known. Assume that we are interested in the distributional shift that replaces the conditional \(q(y|x)\) (an \(N(x, \sigma^2_x)\)-density, evaluated at \(y\)) with an independent \(N(0, \sigma^2_{\epsilon})\) target distribution \(p(y)\). Formally, define the shift factor \(r\) for all \(x, y \in \mathbb{R}\) as

\[ r(x, y) = \frac{p(y)}{q(y|x)}. \]

Then (A2) is satisfied for \(Q\) if and only if

\[ \sigma^2_{\epsilon} < 2(\sigma^2_\epsilon - \sigma^2_X). \]

**Proof.** We begin by directly expanding the second moment of the factor \(r\) under the observational distribution \(Q\) as follows:

\[
\begin{align*}
E_Q [r(X, Y)^2] &= E_Q \left[ \left( \frac{p(Y)}{q(Y|X)} \right)^2 \right] \\
&= \frac{\sigma^2_{\epsilon} \sigma^2_X}{\sigma^2_X} \left[ \exp \left( \frac{(Y - X)^2}{\sigma^2_\epsilon - \sigma^2_X} - \left( \frac{Y}{\sigma} \right)^2 \right) \right] \\
&= \frac{\sigma^2_{\epsilon}}{\sigma^2_X} \left[ \exp \left( \frac{\epsilon^2}{\sigma^2_\epsilon} - \left( \frac{X + \epsilon}{\sigma} \right)^2 \right) \right] \\
&= \frac{\sigma^2_{\epsilon}}{\sigma^2_X} \left[ \exp \left( \left( \frac{\epsilon}{\sigma_\epsilon} \right)^2 - \frac{X^2}{\sigma^2} - \frac{2X \epsilon}{\sigma^2} \right) \right] \\
&= \frac{\sigma^2_{\epsilon}}{\sigma^2_X} \left[ \exp \left( -\frac{X^2}{\sigma^2} - \frac{2 \epsilon \sigma_X}{\sigma^2} \right) \exp \left( \frac{\epsilon^2}{\sigma^2_\epsilon} - \frac{\epsilon^2}{\sigma^2} \right) \right] \\
&= \frac{\sigma^2_{\epsilon}}{\sigma^2_X} \left[ \exp \left( -\frac{\sigma^2_X}{\sigma^2} W + \frac{\epsilon^2}{\sigma^2} \right) \exp \left( \frac{\epsilon^2}{\sigma^2_\epsilon} - \frac{\epsilon^2}{\sigma^2} \right) \right] \\
&= \frac{\sigma^2_{\epsilon}}{\sigma^2_X} \left[ \exp \left( -\frac{\sigma^2_X}{\sigma^2} W + \frac{\epsilon^2}{\sigma^2} \right) \exp \left( \frac{\epsilon^2}{\sigma^2_\epsilon} - \frac{\epsilon^2}{\sigma^2} \right) \right],
\end{align*}
\]

where \(W := \left( \frac{X}{\sigma_X} + \frac{\epsilon}{\sigma_X} \right)^2\). Next, observe that, conditioned on \(\epsilon\), \(W\) has a non-central \(\chi^2(1)\)-distribution with noncentrality parameter \(\frac{\epsilon^2}{\sigma_X^2}\). The moment generating function of \(W\) is given by \(M_W(t) = (1 - 2t)^{-1/2} \exp \left( \frac{\epsilon^2}{\sigma^2_X} - \frac{\epsilon^2}{\sigma^2} \right)\) for all \(t < 1/2\).
Hence, continuing the computation in (I1) and by conditioning on \( \varepsilon \), we get

\[
E_Q \left[ r(X, Y)^2 \right] = \frac{\sigma^2}{\sigma^2} E_Q \left[ \exp \left( -\frac{\sigma^2}{\sigma^2} W \right) \right] = \frac{\sigma^2}{\sigma^2} \left( \frac{1 + 2 \frac{\sigma^2_X}{\sigma^2}}{\left( \frac{\sigma^2}{\sigma^2} \right)^{-2}} \right) E_Q \left[ \exp \left( \frac{\varepsilon^2 - \sigma^2_X}{\sigma^2} \right) 1 + 2 \frac{\sigma^2_X}{\sigma^2} \right] \exp \left( \frac{\varepsilon^2}{\sigma^2} \right)
\]

\[
= \frac{\sigma^2}{\sigma^2} \left( 1 + 2 \frac{\sigma^2_X}{\sigma^2} \right)^{-2} M_S \left( 1 - \frac{\sigma^2}{\sigma^2 + 2\sigma^2_X} \right),
\]

where \( S := (\varepsilon/\sigma)^2 \) and \( M_S \) is the moment generating function of a (central) \( \chi^2 \) distribution. \( M_S(t) \) is finite if and only if \( t < 1/2 \), corresponding to \( 1 - \frac{\sigma^2}{\sigma^2 + 2\sigma^2_X} < 1/2 \) which is equivalent to \( \sigma^2 < 2(\sigma^2_X - \sigma^2) \).

References


