Statistical Testing under Distributional Shifts

Nikolaj Thams*, Sorawit Saengkyongam*, Niklas Pfister*, and Jonas Peters*

University of Copenhagen, Denmark

Abstract

Statistical hypothesis testing is a central problem in empirical inference. Observing data from a distribution P^*, one is interested in the hypothesis $P^* \in H_0$ and requires any test to control the probability of false rejections. In this work, we introduce statistical testing under distributional shifts. We are still interested in a target hypothesis $P^* \in H_0$, but observe data from a distribution Q^* in an observational domain. We assume that P^* is related to Q^* through a known shift τ and formally introduce a framework for hypothesis testing in this setting. We propose a general testing procedure that first resamples from the n observed data points to construct an auxiliary data set (mimicking properties of P^*) and then applies an existing test in the target domain. We prove that this procedure holds pointwise asymptotic level – if the target test holds pointwise asymptotic level, the size of the resample is at most $o(\sqrt{n})$, and the resampling weights are well-behaved. We further show that if the map τ is unknown, it can, under mild conditions, be estimated from data, maintaining level guarantees. Testing under distributional shifts allows us to tackle a diverse set of problems. We argue that it may prove useful in reinforcement learning, we show how it reduces conditional to unconditional independence testing and we provide example applications in causal inference. Code is easy-to-use and will be available online.

1 Introduction

For almost any type of data, a corner stone in statistical method and theory is testing hypotheses about the distribution that generated the data. The framework of testing whether the data generating mechanism, P^*, belongs to a class of distributions H_0 is part of many areas in empirical research, from applied studies to theoretical work.

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 this corresponds to assigning patients to a certain treatment. P^* may also represent a hypothetical distribution of events, which have not yet manifested themselves in reality, as would be the case if P^* represents 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.

In this paper, we assume that the distributional shift $Q \mapsto P$ is known. Even then, it may not always be clear how to utilize the observed data to test the hypothesis $P^* \in H_0$. We propose a general framework to test this hypothesis from observable data Q^*. 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. Our method does not assume full knowledge of Q^* or P^*, but only knowledge of the (potentially unnormalized) ratio p^*/q^*, where q^* and p^* are densities of Q^* and P^* respectively, see (5) below. 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.

*{thams, ss, np, jonas.peters}@math.ku.dk

Our framework resamples the data using a sampling scheme similar to sampling importance resampling (SIR), proposed by Rubin (1987) and Smith and Gelfand (1992). If $X_1, \ldots, X_n \in \mathbb{R}^d$ are sampled i.i.d. from Q^*, our method draws a resample $X_{i_1}, \ldots, X_{i_m}$ of distinct observations, using weights $r(X_i) \propto q^*(X_i)/p^*(X_i)$. We assume existence of a test φ in the target domain, i.e., a test that could be applied if data from P^* were available. We then test the hypothesis $P^* \in H_0$ by applying φ to the resample. We prove that this procedure obtains pointwise asymptotic level, meaning that the probability of type 1 errors of the test $P^* \in H_0$ is asymptotically controlled if φ holds asymptotic level, the weights r have finite second moment in Q^*, and $m = o(\sqrt{n})$. We show, that the same can be obtained, if r is not known, but can be estimated 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 non-linear test statistic.

We showcase different examples, where a problem can be cast as hypothesis test under distributional shifts, including hypothesis tests in off-policy evaluation, tests of conditional independence in an observed distribution and testing the absence of causal edges through dormant independencies (Verma and Pearl, 1991; Shpitser and Pearl, 2008), that is, certain equality constraints in the observed distribution.

Ratios of densities have been applied in the reinforcement learning literature (e.g. Sutton and Barto, 1998), where inference in P^* using data from Q^* is known as off-policy prediction (Precup et al., 2001). Typically, one estimates the expectation of X under P^* using importance sampling, (IS), that is, weighted averages using weights $r(X)$, possibly truncated to decrease variance (Precup et al., 2001; Mahmood et al., 2014). An approach based on resampling was proposed by Schlegel et al. (2019) to predict expectations in P^*, although, they consider the size of the resample fixed and do not consider statistical testing.

In the causal inference literature, inverse probability weighting (IPW) can be used to adjust for confounding or selection bias in data (e.g. Horvitz and Thompson, 1952). To estimate the effect of a treatment X on a response Y, one weighs each observed response Y_i with $1/q^*(X_i|Z_i)$, where Z is an observed confounder. This approach corresponds to our framework of distributional shifts, using a uniform target distribution $p^*(x) \propto 1$. For continuous treatments, some authors change the numerator to a marginal distribution $p^*(x)$ to stabilize the weights (Hernán and Robins, 2006; Naimi et al., 2014). This again coincides with how our framework would be applied for testing hypotheses in causal inference. In general, importance sampling and inverse probability weighting 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 $\mathbb{E}[Y_i]$ or $\mathbb{E}[f(X_i, Z_i)]$, whereas our approach also applies to test statistics that are functions of the entire sample, which is the case for most tests.

SIR sampling schemes were first studied by Rubin (1987) and are often used in the context of Bayesian inference (Smith and 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.

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 $\mathcal{Z} \subseteq \mathbb{R}^d$ 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 $\mathcal{Q}$ over a (potentially) different observational domain $\mathcal{X} \subseteq \mathbb{R}^e$.

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

$$\tau(Q^*) \in H_0. \tag{1}$$

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

$$\sup_{Q \in \tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{2}$$

In practice, requiring level at sample size n is often too restrictive. We say that the test has pointwise asymptotic level α if

$$\sup_{Q \in \tau^{-1}(H_0)} \limsup_{n \to \infty} \mathbb{P}_Q(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{3}$$

Remark. *The map $\tau : \mathcal{Q} \to \mathcal{P}$ 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 τ. Alternatively, one may also start with a map $\eta : \mathcal{P} \to \mathcal{Q}$ and say that the test holds level α at sample size n if*

$$\sup_{P \in H_0} \mathbb{P}_{\eta(P)}(\psi_n(\mathbf{X}_n, U) = 1) \leq \alpha. \tag{4}$$

This corresponds to a level guarantee for a test testing the hypothesis $\eta^{-1}(Q^*) \cap H_0 \neq \emptyset$. If τ 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 $\tau : \mathcal{Q} \to \mathcal{P}$, that is, Equations (1)*,* (2) *and* (3)*.*

2.2 Distributional shifts

In the remainder of this work, we assume that $\mathcal{X}$ and $\mathcal{Z}$ are both subsets of $\mathbb{R}^d$, that is $e = d$, and that all distributions in $\mathcal{P}$ and $\mathcal{Q}$ have densities with respect to the same dominating product measure μ. We refer to a distribution Q and its density q interchangeably.

We consider two types of maps $\tau : \mathcal{Q} \to \mathcal{P}$, 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 $q \in \mathcal{Q}$ the target density $\tau(q)$ satisfies that

$$\tau(q)(x^1, \ldots, x^d) \propto r(x^A) \cdot q(x^1, \ldots, x^d) \qquad \text{for all } (x^1, \ldots, x^d) \in \mathcal{Z}. \tag{5}$$

Here, we assume that the factor r is known in the sense that it can be evaluated for any given x^A. This type of map naturally arises when considering off-policy evaluations with a known training policy or when performing a conditional independence test with known conditional, for example, see Section 3.1.

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

$$\tau(q)(x^1, \ldots, x^d) \propto r_q(x^A) \cdot q(x^1, \ldots, x^d) \qquad \text{for all } (x^1, \ldots, x^d) \in \mathcal{Z}. \tag{6}$$

Here, we assume that the factor $r_{(\cdot)}$ can be evaluated for any given (q, x^A). This case arises, for example, when the training policy or the conditional is unknown and needs to be estimated from data. 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 τ 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 distributions to evaluate $r(x^A)$.

2.3 Exploiting a test in the target domain

In this work, we assume that there is a test φ for the hypothesis H_0 which can be applied to data from the target domain $\mathcal{Z}$. Formally, we consider a sequence $\varphi_m : \mathcal{Z}^m \times \mathbb{R} \to \mathbb{R}$ of (potentially randomized) hypothesis tests for H_0 that can be applied to m observations $\mathbf{Z}_m$ from the target domain $\mathcal{Z}$ and a uniformly distributed random variable V, generating the randomness of φ_m. For simplicity, we omit V from the notation and write $\varphi_m(\mathbf{Z}_m)$. We say that φ_m has pointwise asymptotic level α for H_0 in the target domain if

$$\sup_{P \in H_0} \limsup_{m\to\infty} \mathbb{P}_P(\varphi_m(\mathbf{Z}_m) = 1) \leq \alpha. \tag{7}$$

To address the problem of testing under a distributional shift, we propose to resample a data set of size m from the observed data $\mathbf{X}_n$ (using resampling weights that depend on the shift) and apply the test φ_m to the resampled data (see Section 4 for details). We show that this yields a randomized test ψ satisfying the level requirement (3). This procedure is easy-to-use and can be combined with any testing procedure φ 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 $Q^* \in H_0^{\mathcal{Q}}$ for some $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$. If $\tau(H_0^{\mathcal{Q}}) \subseteq H_0$, any test ψ_n satisfying pointwise asymptotic level (3) for $H_0 \subseteq \mathcal{P}$ can be used as a test for $Q^* \in H_0^{\mathcal{Q}}$, and will still satisfy asymptotic level, see Section 4.2.

Such an approach can be particularly interesting when it is more difficult to test the hypothesis $Q^* \in H_0^{\mathcal{Q}}$ than it is to test $\tau(Q^*) \in H_0$ in the target domain. For example, testing conditional independence in the observed domain can be reduced to (unconditional) independence testing in the target domain and testing a complicated Verma equality (Verma and Pearl, 1991) in the observed distribution can be turned into an independence condition in the target distribution. We now present these and other applications of testing under distributional shifts and provide details of our method and its theoretical guarantees in Section 4.

3 Example applications of testing under distributional shifts

3.1 Off-policy testing

Consider a contextual bandit setup (e.g. Langford and Zhang, 2008; Agarwal et al., 2014). In each round, an agent observes a context $Z := (Z^1, \dots, Z^d)$, and selects an action $A \in \{a_1, \dots, a_L\}$, 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 $\mathbf{X}_n$ of n rounds containing observations $X_i := (Z_i, A_i, R_i)$, $i = 1, \dots, n$. 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

$$H_0 := \{P : \mathbb{E}_P[R] \leq 0 \quad \text{and} \quad p(a|z) = p^*(a|z)\}$$

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 importance sampling or inverse probability weighting can be used, too (see Section 1). But it is also possible to test more involved questions.

Suppose that one of the covariates Z^j is used for selecting actions in $q^*(a|z)$ but not in $p^*(a|z)$. We can then test whether R is independent of Z^j under $p^*(a|z)$. We can use this to test whether R and Z^j are dependent under $q^*(a|z)$, simply because the action A is based on Z^j, or whether Z^j depends on R in other ways, for instance because Z^j has a direct effect on R. This may be relevant for learning sets of features that are invariant across different environments, that is, features Z^J such that $R|Z^J$ is stable across environments. A policy that depends on such invariant features is guaranteed to generalize to unseen environments (Saengkyongam et al., 2021). Another, more involved test for off-policy evaluation can be written as a two-sample test.

3.2 Two-sample testing

We can use the framework to perform a two-sample test, after transforming one of the two samples. To formalize this problem, let us consider the observed distribution q^* over a vector $\mathbf{X} = (X_1, \dots, X_d)$ 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, i.e.,

$$\begin{aligned} q^*(\mathbf{x}|k=1) &\mapsto \quad q^*(\mathbf{x}|k=1) \\ q^*(\mathbf{x}|k=2) &\mapsto \quad \tau(q^*)(\mathbf{x}|k=2). \end{aligned}$$

We can then test whether, after the transformation, the two samples come from the same distribution, i.e., whether, in $\tau(q^*)$,

$$\mathbf{X} \perp\!\!\!\perp K \quad \text{or, equivalently,} \quad q^*(\mathbf{x}|k=1) = \tau(q^*)(\mathbf{x}|k=2)$$

for all $\mathbf{x}$ with positive probability. More concretely, let us assume that in the second sample, we know the conditional $q^*(x_2|x_1, k=2)$ and change it to the known $p^*(x_2|x_1, k=2)$. To formally apply our framework, we then define

$$\begin{aligned} H_0 := \{P :& (X_1, \dots, X_d)_{|K=1} \stackrel{\mathcal{L}}{=} (X_1, \dots, X_d)_{|K=2}, \\ & p(x_1|x_2, k=1) = q^*(x_1|x_2, k=2) \text{ and } p(x_1|x_2, k=2) = p^*(x_1|x_2, k=2)\} \end{aligned}$$

and the shift $\tau(q)(x_1, \dots, x_d, k) := r(x_1, x_2, k) \cdot q(x_1, \dots, x_d, k)$, where

$$r(x_1, x_2, k) := \begin{cases} 1 & \text{if } k = 1 \\ \frac{p^*(x_1|x_2, k=2)}{q^*(x_1|x_2, k=2)} & \text{if } k = 2. \end{cases}$$

Two-sample testing can be used for off-policy evaluation (the setting is described in the previous section). We first 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,

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

With a similar reasoning we can also test, non-parametrically, whether the expected reward under the new policy $p^*(a|z, k=2)$ is larger than under the current policy $q^*(a|z, k=2)$. To do so, we define

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

for example. Section 5.2 shows some empirical evaluations of such tests.

3.3 Conditional independence testing

Let us now 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

$$H_0^{\mathcal{Q}} = \{Q : X \perp\!\!\!\perp Y | Z \text{ and } q(z|x) = q^*(z|x)\}$$

by reducing the problem to an unconditional independence test. The key idea is to factor a density $q \in H_0^{\mathcal{Q}}$ as $q(x, y, z) = q(y|x, z) q^*(z|x) q(x)$, replace[1] the conditional $q^*(z|x)$ by, e.g., a

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

standard normal density $\phi(z)$ to obtain the target density p, and then test for unconditional independence of X and Y.

More formally, we define a corresponding hypothesis in the target domain:

$$H_0^{\mathcal{P}} := \{P : X \perp\!\!\!\perp Y \text{ and } p(z|x) = \phi(z)\}$$

with ϕ being the standard normal density. We can then define a map τ by

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

Considering any $q \in H_0^{\mathcal{Q}}$ 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[2] that $X \perp\!\!\!\perp Y \mid Z$ in q implies $X \perp\!\!\!\perp Y$ in p and therefore $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$. Starting with an independence test φ_m for $H_0^{\mathcal{P}}$, we can thus test $\tau(Q^*) \in H_0^{\mathcal{P}}$, with level guarantee in (3). As we have argued in Section 2.4, this corresponds to testing $Q^* \in H_0^{\mathcal{Q}}$, and thereby reduces the question of conditional independence to independence.

If, instead of $q^*(z|x)$, we know the conditional $q^*(x|z)$, we can use the same reasoning as above using the factorization $q(x,y,z) = q(z)q^*(x|z)q(y|x,z)$ and a marginal target density $\phi(x)$ to again test $X \perp\!\!\!\perp Y \mid Z$. When X is a treatment, Y the outcome, Z is the full set of covariates, and $q^*(x|z)$ represents the randomization scheme, this corresponds to testing (non-parametrically) the existence of a causal effect (e.g., Pearl, 2009; Hernán and Robins, 2020; 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 $H_0^{\mathcal{Q}} := \{Q : X \perp\!\!\!\perp Y|Z\}$, $H_0^{\mathcal{P}} := \{P : X \perp\!\!\!\perp Y \text{ and } p(z|x) = \phi(z)\}$, and the map τ via $\tau(q)(x,y,z) := \frac{\phi(z)}{q(z|x)} \cdot q(x,y,z)$, for all $(x^1,\dots,x^d) \in \mathcal{Z}$; 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. In general, such an approach may have less power than more targeted conditional independence tests but is an interesting alternative if we can estimate one of the conditionals well, e.g., because there are many more observations of (X,Z) than there are of (X,Z,Y).

The assumption of knowing one conditional $q(x|z)$ is also exploited by the conditional randomization (CRT) and the conditional permutation test (CPT) by Candès et al. (2018) and Berrett et al. (2020), respectively. 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 cannot only exploit knowledge of $q^*(x|z)$ but also knowledge of $q^*(z|x)$.

3.4 Dormant independences

Let us consider a random vector $(X^1,\dots,X^d)$ 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; Pearl, 2009), we then have for all disjoint subsets $A,B,C \subset \{1,\dots,d\}$ that $X^A \perp\!\!\!\perp X^B \mid X^C$ if A d-separates $B \mid C$. (Whether a d-separation statement holds is entirely determined from the graph; the precise definition of d-separation (e.g., Spirtes et al., 2000) is not important here.) 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 1 shows a

[2]The following statement holds because, clearly, $p(z|x) = \phi(z)$ and if $X \perp\!\!\!\perp Y \mid Z$ in q, that is, $q(y|x,z) = q(y|z)$ for all x,y,z yielding this expression well-defined, it follows $p(x,y) = p(x)p(y)$.

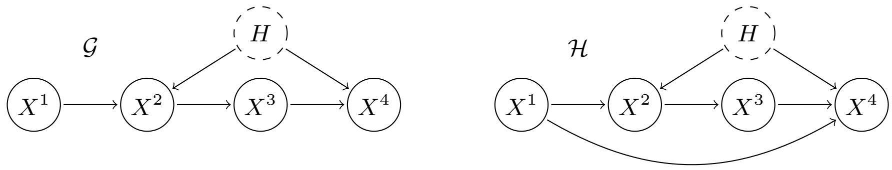


Figure 1: If Q is Markovian w.r.t. graph $\mathcal{G}$ (left), then Q satisfies the Verma constraint (8). In general, this constraint does not hold if Q is Markovian w.r.t. $\mathcal{H}$ (right). Such constraint can be tested for using the framework of statistical testing under distributional shifts, see Section 3.4.

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 1 (left), there exists a function f such that

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

(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 1, right). Assume that the conditional $q(x^3|x^2) = q^*(x^3|x^2)$ is known (e.g., through a randomization experiment). We can then hope to test for this constraint (allowing us to distinguish between $\mathcal{G}$ and $\mathcal{H}$) and define

$$H_0^{\mathcal{Q}} := \{Q : Q \text{ satisfies (8) and } q(x^3|x^2) = q^*(x^3|x^2)\}.$$

Constraints of the form (8) have been studied recently, and in a few special cases, such as binary data, the constraints can be exploited to construct score-based structure learning methodology (Shpitser et al., 2012). 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 Shpitser et al., 2014; Richardson et al., 2017), and Shpitser et al. (2009) proposes 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

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

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 5.4 for a simulation study. In special cases, such as binary data, applying resampling methodology to this type of problem has been considered before (Bhattacharya, 2019), but we are not aware of any work proposing a general testing procedure with theoretical guarantees.

4 Testing by Resampling

4.1 A consistent statistical test based on resampling

We propose a method for testing the target hypothesis $\tau(Q^*) \in H_0$, see (1), using a sample $\mathbf{X}_n$ from the observational distribution Q^*. The idea is to take an existing hypothesis test φ_m for the hypothesis H_0 in the target domain and apply it to a resampled version of the observed data, which mimics a sample in the target domain. We will see that, under suitable assumptions and whenever the original test φ_m holds pointwise asymptotic level in the target

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

Input: Data $\mathbf{X}_n$, target sample size m, hypothesis test φ_m, shift factor $r(x^A)$.

1: $(i_1, \ldots, i_m) \leftarrow$ sample from $\{1, \ldots, n\}^m$ with weights (10) (see Appendix C)

2: $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U) \leftarrow (X_{i_1}, \ldots, X_{i_m})$

return $\psi^r_n(\mathbf{X}_n, U) \coloneqq \varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$

domain, see (7), the overall procedure ψ^r_n has pointwise asymptotic level for the null hypothesis $\tau(Q^*) \in H_0$, see (3).

We consider the setting, where $\tau(q)(x) \propto r(x)q(x)$ for a known shift r; see (5). First, we draw a weighted resample of size m from $\mathbf{X}_n$ similar to the sampling importance resampling (SIR) scheme proposed by Rubin (1987) but using a sampling scheme DRPL ('distinct replacment') that is different from sampling with or without replacement. More precisely, we draw a resample $(X_{i_1}, \ldots, X_{i_m})$ from $X_1, \ldots, X_n$ where $(i_1, \ldots, i_m) \in \{1, \ldots, n\}^m$ is a sequence of distinct[3] values; the probability of drawing the sequence $(X_{i_1}, \ldots, X_{i_m})$ is

$$w_{(i_1,\ldots,i_m)} \propto \begin{cases} \prod_{\ell=1}^m r(X_{i_\ell}) \propto \prod_{\ell=1}^m \frac{\tau(q)(X_{i_\ell})}{q(X_{i_\ell})} & \text{if } (i_1, \ldots, i_m) \text{ is distinct and} \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

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 $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)$, where U is a random variable representing the randomness of the resample. When m is fixed and n approaches infinity, the target sample $\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)$ converges in distribution to m i.i.d. draws from the target distribution $\tau(Q^*)$; see Skare et al. (2003) for a proof based on a slightly different sampling scheme.

Second, we apply the test φ_m for the target domain to the resampled data to obtain a hypothesis test ψ^r_n on the observed data for the target hypothesis (1), $\psi^r_n(\mathbf{X}_n, U) \coloneqq \varphi_m(\Psi^{r,m}(\mathbf{X}_n, U))$; see Algorithm 1.

4.2 Theoretical guarantees

We now prove that the hypothesis test ψ^r_n introduced in the previous section satisfies pointwise asymptotic level. To do so, we require three assumptions: the target test φ_m needs to satisfy pointwise asymptotic level, m and n have to approach ∞ at a suitable rate, and we require the weights to be well-behaved. More precisely, we will make the following assumptions.

(A1) The sequence of hypothesis tests $(\varphi_m)_m$ in the target domain has pointwise asymptotic level α, see (7).

(A2) $m = o(\sqrt{n})$ for $n \to \infty$.

(A3) For all $Q \in \tau^{-1}(H_0)$, it holds that $\mathbb{E}_Q[r(X_i)^2] < \infty$.

(A1) states that if data from the target distribution were available, we would have a valid test for the target hypothesis – if this was not the case, we could hardly expect to be able to test the target hypothesis from observed data. (A2) states that m and n may not approach ∞ in an arbitrary way. (A3) is a condition to ensure the weights are sufficiently well behaved. If $r(x^A)$ only depends on a subset A of variables, and x^A takes finitely many values, (A3) is trivially satisfied. In the case of an off-policy hypothesis test, such as the one described in Section 3.1, a sufficient but not necessary condition is that the policy $q^*(a|z)$ is randomized, 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 $q(x^j|x^{j'})$ to a Gaussian marginal $p(x^j)$, we show in Appendix E under which conditions (A3) is satisfied. If the hypothesis of interest is in the observation domain,

[3] We use 'distinct' and 'non-distinct' 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 original sample $\mathbf{X}_n$.

corresponding to the setting described in Section 2.4, we are usually free to chose any target density, so we can ensure that the tails decay sufficiently fast to satisfy (A3). In Section 5.1 below, we validate (A2) and (A3) in synthetic data, and find the assumptions are empirically relevant for whether our test has the correct level or not. We now state the first main result. All proofs can be found in Appendix D.

Theorem 1 (known weights). *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. 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). Given a sequence of tests φ_m, let $\psi_n^r(\mathbf{X}_n, U) := \varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$ denote the output of Algorithm 1, and let $\alpha \in (0, 1)$. Then if (A1), (A2) and (A3) are true, it holds that*

$$\sup_{Q \in \tau^{-1}(H_0)} \limsup_{n\to\infty} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \le \alpha,$$

i.e., ψ_n^r satisfies pointwise asymptotic level α for the hypothesis $\tau(Q^*) \in H_0$.

Theorem 1 considers the case in which the known shift factor r does not depend on q. Next, we consider the setting in which the shift factor is allowed to explicitly depend on q. This is relevant, for example, if the shift τ represents a change of the conditional of a variable X^j from $q^*(x^j|x^B)$ to $p^*(x^j|x^C)$, say, but the observational conditional $q^*(x|x^B)$ is unknown, corresponding to the setting in (6). If $q^*(x|x^B)$ is unknown, we are not able to compute the weights $w_i \propto r(X_i) \propto p^*(X_i^j|X_i^B)/q^*(X_i^j|X_i^C)$. However, we can still try to estimate $q^*(x^j|x^B)$ and obtain approximate weights $\hat{r} \propto p^*/\hat{q}^*$. Then, if we make the following modifications to (A2) and (A3),

(A2') $m = o(\min(n^a, n^{1/2}))$ for $n \to \infty$,

(A3') For all $Q \in \tau^{-1}(H_0)$ it holds that $E_Q[r_q(X_i)^2] < \infty$,

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 2 (estimated weights). *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift, satisfying $\tau(q)(x) \propto r_q(x)q(x)$, see (6). Let $\hat{r}_n$ be an estimator for r_q such that there exists $a \in (0, 1)$ satisfying for all $Q \in \tau^{-1}(H_0)$ that*

$$\lim_{n\to\infty} \sup_{x \in \mathcal{X}} \mathbb{E}_Q \left| \left(\frac{\hat{r}_n(x)}{r_q(x)} \right)^{n^a} - 1 \right| = 0,$$

where the expectation is taken over the randomness of $\hat{r}_n$. Partition $\mathbf{X}_n$ into two data sets $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ such that $n_1^a = \sqrt{n_2}$. Estimate $\hat{r}_{n_1}$ using $\mathbf{X}_{n_1}$ and apply Algorithm 1 to $\mathbf{X}_{n_2}$ using the estimated weights $\hat{r}$. Given a sequence of tests φ_m, let $\psi_n^{\hat{r}} := \varphi(\Psi^{\hat{r}_{n_1},m}(\mathbf{X}_{n_2}, U))$ be the output of the combined procedure, as detailed in Algorithm 2 in Appendix B, and let $\alpha \in (0, 1)$. Then, if (A1), (A2') and (A3') are true, it holds that

$$\sup_{Q \in \tau^{-1}(H_0)} \limsup_{n\to\infty} \mathbb{P}_Q(\psi_n^{\hat{r}}(\mathbf{X}_n, U) = 1) \le \alpha,$$

i.e., $\psi_n^{\hat{r}}$ satisfies pointwise asymptotic level α for the hypothesis $\tau(Q^*) \in H_0$.

In Section 2.4, we argued that one can use the framework to test a hypothesis in the observational domain, too. Indeed, the above results directly imply the following corollary (see Corollary 3 in Appendix D.4 for a more detailed version).

Corollary 1 (Hypothesis testing in the observational domain). *Consider hypotheses $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$ and $H_0^{\mathcal{P}} \subseteq \mathcal{P}$ and let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift such that $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$. Under the same assumptions as in Theorem 1, ψ_n^r satisfies pointwise asymptotic level α for the hypothesis $Q^* \in H_0^{\mathcal{Q}}$.*

4.3 Resampling Schemes

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

To do so, let $\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)$ and $\Psi^{r,m}_{\texttt{NO-REPL}}(\mathbf{X}_n, U)$ denote weighted sampling with and without replacement, respectively, both of which are implemented in most standard statistical software packages. Though $\Psi_{\texttt{DRPL}}$ and $\Psi_{\texttt{NO-REPL}}$ 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 C). We can sample from $\Psi_{\texttt{DRPL}}$ by sampling from $\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n)$ and rejecting the sample until the indices $(i_1, \ldots, i_m)$ are distinct, see Appendix C.1. In Proposition 1 (Appendix C) we prove that under suitable assumptions, such as $m = o(\sqrt{n})$, the probability of drawing a distinct sequence already in a single draw $\Psi_{\texttt{REPL}}$ approaches 1, when $n \to \infty$.

In some cases (though these typically only occur when (A2) or (A3) are violated, where our level result does not apply), the above rejection sampling from $\Psi_{\texttt{REPL}}$ may take a long time to accept a sample. For these cases, we propose to use an (exact) rejection sampler based on $\Psi_{\texttt{NO-REPL}}$, which will typically be faster (since it has the same support as $\Psi_{\texttt{DRPL}}$). We provide all details in Appendix C.2.

If neither of the two exact sampling schemes for $\Psi_{\texttt{DRPL}}$ is computationally feasible, we provide an approximate sampling method that applies a Gibbs sampler to a sample from $\Psi_{\texttt{NO-REPL}}$; we refer to this scheme as $\Psi_{\texttt{DRPL-GIBBS}}$. The details are provided in Appendix C.3.

The above mentioned result that the probability of drawing distinct observations in $\Psi_{\texttt{REPL}}$ approaches 1, when $n \to \infty$, has another implication. We prove that we can obtain the same level guarantee, when using $\Psi_{\texttt{REPL}}$ instead of $\Psi_{\texttt{DRPL}}$ (see Corollary 2 in Appendix C). This result, however, requires an assumption that is stronger than (A3). Intuitively, stronger assumptions are required for $\Psi_{\texttt{REPL}}$ because sampling with replacement is much more prone to experience large variance due to single observations with enormous weights.

5 Experiments

5.1 Exploring assumptions (A2) and (A3)

We explore the impact of violating either (A2), stating that $m = o(\sqrt{n})$, or (A3), stating that the weights must have finite second moment in the observational distribution. To do so, we apply the procedure discussed in Section 3.3 that reduces a conditional independence test $X \perp\!\!\!\perp Y \mid Z$ in the observational domain to an unconditional independence test in the target domain. Specifically, we simulate $n \in \{100, 300\}$ i.i.d. observations from the linear Gaussian model with

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

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 $\phi_\sigma(z)$ with mean zero and variance σ^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 $\theta = 2$ (where, ideally, we reject the hypothesis) and for $\theta = 0$ (where, ideally, we accept the hypothesis). The procedure of simulating, resampling and testing (at level $\alpha = 0.05$) is repeated $5'000$ times, and the resulting rejection rates of the test are plotted in Figure 2.

First, we test the importance of (A2) by changing the rate c for the size $m = \lfloor n^c \rfloor$ of the resampled data set. Sampling from $\Psi_{\texttt{DRPL}}$ becomes difficult when assumption (A2) is substantially violated, so we also sample from $\Psi_{\texttt{DRPL-GIBBS}}$ (see Appendix C), which is an approximation of $\Psi_{\texttt{DRPL}}$, and sample with replacement, $\Psi_{\texttt{REPL}}$. As suggested by the theory, $\Psi_{\texttt{DRPL}}$ holds level, when $c = 0.5$ (blue line). In fact, the threshold of $c = 0.5$ seems too restrictive for $\Psi_{\texttt{DRPL}}$, and the level only deviates from 5% at around $c = 0.7$ for the approximate $\Psi_{\texttt{DRPL-GIBBS}}$. While $\Psi_{\texttt{REPL}}$ deviates

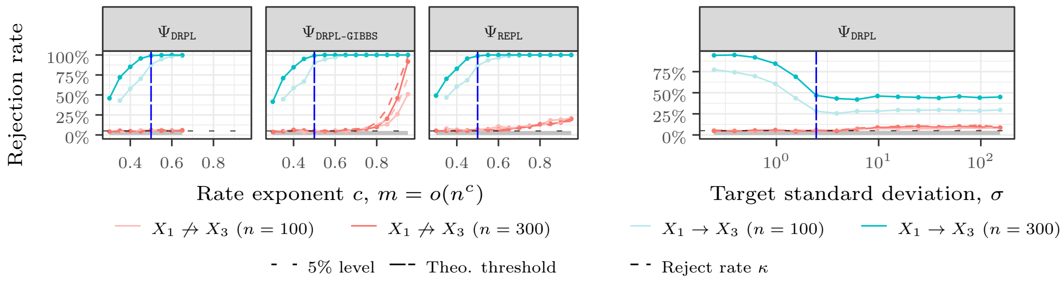


Figure 2: *(Both)* In both experiments, we replace the conditional distribution $q^*(z|x)$ with a marginal distribution $\phi_\sigma(z)$. We perform a test for independence $X \perp\!\!\!\perp Y$ in the target distribution, and plot the rejection rates. We also plot the rejection rates of a test κ (dashed lines) which tests whether the empirical conditional distribution of $Z \mid X$ in the resample $\Psi^{r,m}(\mathbf{X}_n)$ is significantly different from the target density p. *(Left)* Validation of (A2). Our procedure is run with $m = o(n^c)$ for different exponents c. As suggested by the theory, the level holds if (A2) is satisfied (left of the blue vertical line). The level deteriorates when κ rejects. $\Psi_{\texttt{DRPL}}$ is only run up to $c \le 0.65$, because sampling becomes computationally challenging if (A2) is further violated. *(Right)* Validation of (A3). Our procedure is run with different standard deviations σ in the Gaussian target distribution $\phi_\sigma(z)$. The blue vertical line indicates the theoretical threshold of $\sqrt{6}$, see Section 5.1.

from the 5% level earlier than $\Psi_{\texttt{DRPL-GIBBS}}$, it is less adversarially effected than $\Psi_{\texttt{DRPL-GIBBS}}$ when c approaches 1, because as c approaches 1, the entire observed sample is drawn by $\Psi_{\texttt{DRPL-GIBBS}}$, and the test is performed on the observed data.

Second, we test the importance of (A3). We show in Appendix E that (A3) is satisfied for a Gaussian conditional specified by $Z = X + \varepsilon_Z$ if and only if $\sigma^2 < 2(\sigma^2_{\varepsilon_Z} - \sigma^2_X)$, where $X \sim \mathcal{N}(0, \sigma^2_X)$ and $\varepsilon_Z \sim \mathcal{N}(0, \sigma^2_{\varepsilon_Z})$. In this experiment, it follows that (A3) holds if and only if $\sigma < \sqrt{6}$. We observe that when σ exceeds the threshold of $\sqrt{6}$ (blue line), the level eventually deviates from the 5% level. Further, the power drops when σ approaches the threshold.

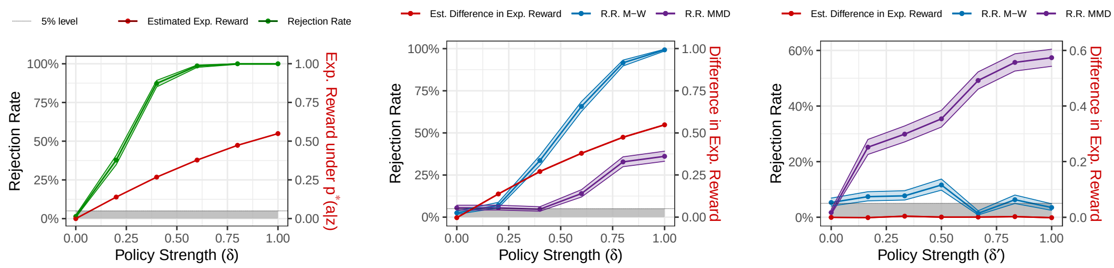
In a finite sample setting, it is not straightforward to verify the assumptions (A2) and (A3) directly. However, we can test whether the resampling worked as intended before testing our hypothesis. Given that the shift corresponds to changing $q^*(x^j|x^C)$ to $p^*(x^j|x^B)$, we can verify whether the conditional $X^j \mid X^B$ in the resampled data $\Psi^{r,m}(\mathbf{X}_n)$ is close to the target conditional $p^*(x^j|x^B)$, e.g., by simulating new data from $p^*(x^j|x^B)$ and conducting a two-sample test. When considering a marginal distribution $p^*(x^j)$, as we do in this experiment, we can also validate the assumptions by testing whether Z is marginally independent of X in the resampled data $\Psi^{r,m}(\mathbf{X}_n)$. Since the data are Gaussian, we can do so by a simple correlation test κ. The corresponding rejection rates of κ are also given in Figure 2. We can see that this empirical verification of the assumptions works well and allows to empirically justify the assumptions.

5.2 Off-policy testing

We apply our method to perform statistical testing in an off-policy contextual bandit setting as discussed in Section 3.1. We generate a data set $\mathbf{X}_n$, ($n = 30'000$), consisting of observations $X_i = (Z_i, A_i, R_i)$ drawn according to the following data generating process:

$$Z := \varepsilon_Z \qquad A \mid Z \sim q^*(A \mid Z) \qquad R := \beta_A^\top Z + \varepsilon_R,$$

where $\varepsilon_Z \sim \mathcal{N}(0, I_3)$ and $\varepsilon_R = \mathcal{N}(0, 1)$, A takes values in the action space $\{a_1, \ldots, a_L\}$, $q^*(a|z)$ denotes an initial policy that was used to generate the data $\mathbf{X}_n$ 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 $z \in \mathbb{R}^3$, $q^*(a|z) = 1/L$.



(a) (b) (c)

Figure 3: Off-policy statistical testing: (a) one-sample test for testing whether the mean under $p^*_\delta(a|z)$ is less or equal to 0, see Section 3.1. (b), (c) two-sample tests for testing whether the reward under $p^*_\delta(a|z)$ and $p^*_{\delta'}(a|z)$ respectively has a different distribution than the reward under the initial policy, see Section 3.2. In all cases, the null hypothesis is true for $\delta = 0$. The target policies affect the mean of the reward in (a) and (b), whereas they affect its variance in (c).

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 two hypotheses, namely difference in means and difference in distributions. Specifically, in the first experiment, we construct different target policies $p^*_\delta(a|z)$ with increasing δ puts more mass on the optimal action (and thereby deviates from the initial policy). More precisely, $p^*_\delta(a|z)$ is a linear softmax policy, i.e., $p^*_\delta(a|z) \propto \exp(\delta \beta_a^\top z)$. When $\delta = 0$, the target policy reduces to a uniform random policy. As $\delta \to \infty$, the target policy converges to an optimal policy. We then apply our method to non-parametrically test whether $\mathbb{E}_{P^*_\delta}(R) \leq 0$ on the target distribution in which the policy $p^*_\delta(a|z)$ is used. We employ the Wilcoxon signed-rank test (Wilcoxon, 1992) in the target domain. For $\delta = 0$, the expected reward is zero and for increasing δ the expected reward increases, too. Figure 3(a) shows that for $\delta = 0$, our method indeed holds the correct level and starts to correctly reject for increasing δ. For comparison, we include an estimate of the expected reward based on inverse probability weighting.

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 3.2 to test whether $R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2}$, where $K = 1$ indicates a sample under the initial policy and $K = 2$ indicates a sample under a target policy. We consider two non-parametric tests, namely a kernel two-sample test based on the maximum mean discrepancy (MMD) (Gretton et al., 2012) and the Mann-Whitney (M-W) U test (Mann and Whitney, 1947). Here, for $\delta = 0$, the two policies coincide and for $\delta > 0$, there is a difference in the expected reward. Indeed, both tests are able to detect this difference. The M-W U performs slightly better, see Figure 3(b).

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., $p^*_{\delta'}(a|z) \propto \delta'$ 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 δ' is 1, the target policy is the same as the initial policy, whereas when δ' increases, the variance of the reward becomes smaller (in Figure 3(c), δ' is rescaled to 0–1 range). We then apply the same two-sample testing methods used in the second experiment to test whether $R_{|K=1} \overset{\mathcal{L}}{=} R_{|K=2}$. This time, the MMD has more power, see Figure 3(c). In all three experiments, we set the false positive rate to 5%.

5.3 Testing a conditional independence with a complex conditional

In the setting of conditional independence testing, we now compare our method – which turns the problem into a test for unconditional independence as discussed in Section 3.3 – to existing conditional independence tests. We sample $n = 150$ observations from the following structural

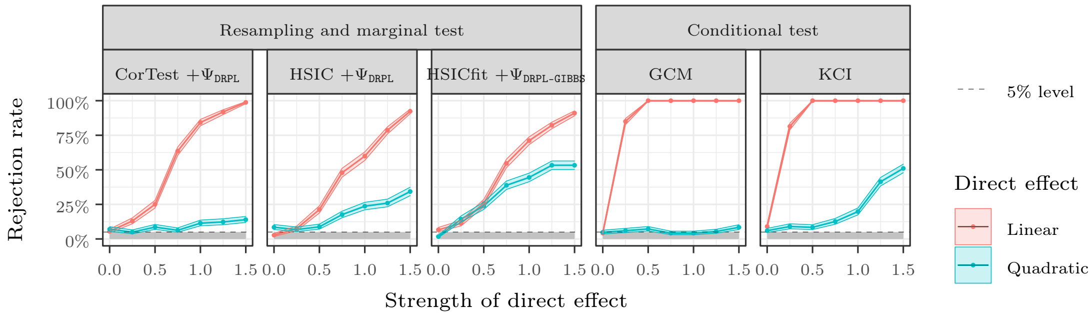


Figure 4: Rejection rates for conditional independence tests $X \perp\!\!\!\perp Y \mid Z$ in the setting from (11). The direct effect of X on Y is θX^τ, where the exponent τ is either 1 (red) or 2 (blue), and we plot θ on the x-axis. The left panel shows our method of resampling and testing marginal independence, using either HSIC or a Pearson correlation test to test independence. We compare to two conditional independence tests, GCM and KCI.

causal model

$$X := \texttt{GaussianMixture}(-2, 2) \quad Z := -X^2 + \varepsilon_Z \quad Y := \sin(Z) + \theta X^\tau + \varepsilon_Y, \tag{11}$$

inducing a distribution Q^*, where `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 $\mathcal{N}(0, 4)$-variables and $\theta \in [0, 2]$, $\tau \in \{1, 2\}$ are hyper-parameters. Considering the conditional $q^*(z|x)$ to be known, we apply our methodology for testing conditional independence $X \perp\!\!\!\perp Y \mid Z$ using $m = \sqrt{n}$. (As the model above is an additive noise model with non-linear functions, the presence or absence of a causal effect from X to Y is in principle identifiable from data (Hoyer et al., 2008; Peters et al., 2014), so such methods could in principle be used, too.) 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 generalized covariance measure (GCM) by Shah and Peters (2020) and a kernel conditional independence (KCI) by from Zhang et al. (2011). 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 generalized additive model. Here, HSICfit is used with the approximate $\Psi_{\texttt{DRPL-GIBBS}}$.

We repeat the experiment $1'000$ times and plot the rejection rates in Figure 4 at various strengths θ of the edge $X \to Y$. All instances of our method have the correct level, see rejection rates for $\theta = 0$. When $\tau = 1$, i.e., the direct effect $X \to Y$ 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 do not have any power, as expected since they are based on correlations. KCI and HSIC have comparable power in the quadratic case, with our approach even obtaining slightly more power than KCI for large θ.

This validates the proposal in Section 3.3, that our method of resampling and testing marginal independence can be used as a conditional independence test. 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 (A2) is chosen much smaller than n.

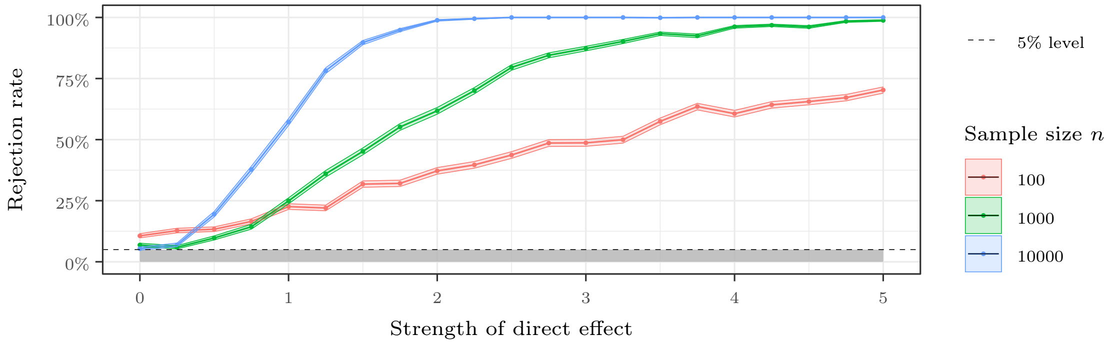


Figure 5: Rejection rates for testing a dormant independence. We test the dormant independence discussed in Section 3.4, where the existence of a direct causal link $X_1 \to X_4$ in Figure 1 cannot be identified by a conditional independence statement. Instead, the effect $X_1 \to X_4$ is identified by a Verma constraint in the observational distribution, which can be tested as an independence statement in the target distribution. We vary the strength θ of the causal link $X_1 \to X_4$. For $\theta = 0$, ideally the test rejects the target independence in 5% of cases, and for $\theta > 0$ ideally in 100% of cases.

5.4 Verma graph

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 factorizes according to the graph $\mathcal{H}$ in Figure 1 and test the existence of the edge $X_1 \to X_4$. 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 in the target distribution $Q^{\mathrm{do}(X^3 := N)}$, which can be obtained by applying our method using $r_q(x^3, x^2) := q^{\mathrm{do}}(x^3)/q(x^3|x^2)$.

More precisely, we simulate a system, where H, X^1, and X^2 are Gaussian and X^3 and X^4 are binary random variables. We estimate the conditional $q(x^3|x^2)$ by logistic regression and use a target distribution specified by $p(x^3) = 1/2$ for $x^3 \in \{0, 1\}$. We then test for independence of X^1 and X^4 in the interventional distribution using an HSIC test (Gretton et al., 2008). Starting from $\theta = 0$ (where the target hypothesis is true), we iterate the strength of the causal effect θ.

The resulting rejection rates, for several n, are shown in Figure 5. We observe that our method identifies both the absence and presence of the causal edge $X_1 \to X_4$, although the power only increases slowly for small sample sizes indicating the statistical challenge that comes with testing a Verma constraint.

6 Conclusion

We formally introduce statistical testing under distributional shifts and illustrate that it can be applied in diverse areas such as reinforcement learning, 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 independencies – for which to-date no test with provable level guarantees exists – and it performs remarkably well both on complicated conditional independence and off-policy testing problems. Its key strength is that it is extremely easy to apply and can be combined with any existing test making it an attractive go-to method for complicated testing problems.

There are several areas which we believe are important and merit further investigation. Firstly, a limiting factor of our procedure is its reliance on rather large sample sizes. This is also shown by our theory, which assumes the size of the resample m to be at most $o(\sqrt{n})$. While it may be difficult to improve this result for arbitrary target tests, the experiments suggest that, particularly in the finite sample regime, it might be possible to provide ways of choosing m larger without breaking the level guarantees. Secondly, it would be interesting to extend the theoretical results further both in terms of power and by finding sufficient conditions for achieving uniform asymptotic level. Thirdly, better theoretical understanding of the differences between sampling methods, including sampling with replacement, and of whether drawing distinct samples can alleviate the problems of outlier weights in importance sampling, for example, could help us to find the optimal resampling scheme for the problem at hand. Finally, we believe that statistical testing under distributional shifts has the potential of yielding testing procedures in a wide area of applications. For a given application, it may happen that the resampling methods that we propose, have less power than other, more specialized procedures, but they are easy to understand and simple to use. We hope that seeing more of such applications may shed further light on the question of how to characterize the types of distributional shifts we see in real world applications.

Acknowledgments and Disclosure of Funding

We thank Peter Rasmussen for helpful discussions about the combinatorics. NT, SS, and JP were supported by a research grant (18968) from VILLUM FONDEN and JP was, in addition, supported by the Carlsberg Foundation.

References

A. Agarwal, D. Hsu, S. Kale, J. Langford, L. Li, and R. Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *Proceedings of the 31st International Conference on Machine Learning (ICML)*, pages 1638–1646. PMLR, 2014.

J. Aldrich. Autonomy. *Oxford Economic Papers*, 41:15–34, 1989.

T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(1):175–197, 2020.

R. Bhattacharya. On actually testing generalized independence (verma) constraints, 2019. `https://www.cs.jhu.edu/~rohit/posts/verma.pdf`, last accessed 15.04.2021.

S. Bongers, P. Forre, J. Peters, and J. M. Mooij. Foundations of structural causal models with cycles and latent variables. *Annals of Statistics (accepted), ArXiv e-prints (1611.06221v5)*, 2021.

E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741, 1984.

A. Gretton, K. Fukumizu, C. Teo, L. Song, B. Schölkopf, and A. Smola. A kernel statistical test of independence. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

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

T. Haavelmo. The probability approach in econometrics. *Econometrica*, 12:S1–S115 (supplement), 1944.

M. A. Hernán and J. M. Robins. Estimating causal effects from epidemiological data. *Journal of Epidemiology & Community Health*, 60(7):578–586, 2006.

M. A. Hernán and J. M. Robins. *Causal inference: What if.* Boca Raton: Chapman & Hall/CRC, 2020.

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

P. Hoyer, D. Janzing, J. M. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21, pages 689–696, 2008.

J. Langford and T. Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

S. Lauritzen. *Graphical Models.* Oxford University Press, New York, NY, USA, 1996.

S. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer. Independence properties of directed Markov fields. *Networks*, 20(5):491–505, 1990.

A. R. Mahmood, H. P. van Hasselt, and R. S. Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

A. I. Naimi, E. E. Moodie, N. Auger, and J. S. Kaufman. Constructing inverse probability weights for continuous exposures: a comparison of methods. *Epidemiology*, pages 292–299, 2014.

J. Pearl. *Causality.* Cambridge University Press, 2009.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15:2009–2053, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: Foundations and learning algorithms.* The MIT Press, 2017.

D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the 18th International Conference on Machine Learning (ICML)*, pages 417–424, 2001.

T. S. Richardson, R. J. Evans, J. M. Robins, and I. Shpitser. Nested Markov properties for acyclic directed mixed graphs. *ArXiv e-prints (1701.06686)*, 2017.

D. B. Rubin. The calculation of posterior distributions by data augmentation: Comment: A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: The sir algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987.

S. Saengkyongam, N. Thams, J. Peters, and N. Pfister. Invariant policy learning: A causal perspective. *(under preparation)*, 2021.

M. Schlegel, W. Chung, D. Graves, J. Qian, and M. White. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

R. Serfling. *Approximation theorems of mathematical statistics.* Wiley series in probability and mathematical statistics : Probability and mathematical statistics. Wiley, 1980.

R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

I. Shpitser and J. Pearl. Dormant independence. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pages 1081–1087. AAAI Press, 2008.

I. Shpitser, T. Richardson, and J. Robins. Testing edges by truncations. In *IJCAI-09 - Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI International Joint Conference on Artificial Intelligence, pages 1957–1963. International Joint Conferences on Artificial Intelligence, Jan. 2009. ISBN 9781577354260. 21st International Joint Conference on Artificial Intelligence, IJCAI 2009 ; Conference date: 11-07-2009 Through 16-07-2009.

I. Shpitser, T. S. Richardson, J. M. Robins, and R. Evans. Parameter and structure learning in nested Markov models. *ArXiv e-prints (1207.5058)*, 2012.

I. Shpitser, R. J. Evans, T. S. Richardson, and J. M. Robins. Introduction to nested Markov models. *Behaviormetrika*, 41:3–39, 2014.

Ø. Skare, E. Bølviken, and L. Holden. Improved sampling-importance resampling and reduced bias importance sampling. *Scandinavian Journal of Statistics*, 30(4):719–737, 2003.

A. F. Smith and A. E. Gelfand. Bayesian statistics without tears: A sampling–resampling perspective. *The American Statistician*, 46(2):84–88, 1992.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings of the 6th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 255–270, 1991.

F. Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

K. Zhang, J. Peters, D. Janzing, and B. Schölkopf. Kernel-based conditional independence test and application in causal discovery. In *Proceedings of the 27th Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 804–813. AUAI Press, 2011.

A Further comments on the framework

A.1 Forward and backward shifts, τ and η

In this paper, as mentioned in Section 2.1, we mostly 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 η such that $Q^* = \eta(P^*)$ is a distribution which can be sampled from in practice. If τ and η 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 $Q \in \{Q \mid \eta^{-1}(Q) \cap H_0 \neq \emptyset\}$ (η view) or over $Q \in \{Q | \tau(Q) \in H_0\}$ (τ view). To see this, we first start with the (natural) level guarantee from the η view: $\sup_{P\in H_0} \mathbb{P}_{\eta(P)}(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha$. We then have

$$
\begin{aligned}
& \sup_{P\in H_0} \mathbb{P}_{\eta(P)}(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha \\
\Leftrightarrow \quad & \sup_{Q\in\eta(H_0)} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha \\
\Leftrightarrow \quad & \sup_{Q\in\{Q' \mid \eta^{-1}(Q')\cap H_0 \neq\emptyset\}} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha.
\end{aligned}
$$

If, alternatively, we start with the level guarantee from the τ view, we find

$$
\begin{aligned}
& \sup_{P\in H_0} \sup_{Q\in\tau^{-1}(P)} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha \\
\Leftrightarrow \quad & \sup_{Q\in\tau^{-1}(H_0)} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha \\
\Leftrightarrow \quad & \sup_{Q\in\{Q' \mid \tau(Q')\in H_0\}} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha.
\end{aligned}
$$

Comparing the last two lines yields the claim.

A.2 Example: Interventions in causal models

One example of a distributional shift τ is the case where τ represents an intervention in a structural causal model (SCM) over $X^1, \ldots, X^d$ (Pearl, 2009). An SCM $\mathcal{S}$ over $X^1, \ldots, X^d$ is a collection of structural assignments $f^1, \ldots, f^d$ and noise distributions $Q_{N^1}, \ldots, Q_{N^1}$ such that for each $j = 1, \ldots, d$, we have $X^j := f^j(\mathrm{PA}^j, N^j)$. Here, the noise variables N^j are distributed according to $N^j \sim Q_{N^j}$ and are assumed to be jointly independent. The sets $\mathrm{PA}^j \subseteq \{X^1, \ldots, X^d\} \backslash \{X^j\}$ 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 $\mathcal{S}$ 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 factorization property (see Lauritzen et al. (1990) or Theorem 1.4.1 in Pearl (2009)): $q(x^1, \ldots, x^d) = \prod_{j=1}^d q_{X^j|\mathrm{PA}^j}(x^j \mid x^{\mathrm{PA}^j})$. In an SCM, an intervention on a variable X^k replaces the tuple $(f^k, \mathrm{PA}^k, Q_{N^k})$ with $(\bar{f}^k, \bar{\mathrm{PA}}^k, \bar{Q}_{N^j})$ in the structural assignment for X^k, and we denote the replacement by $\mathrm{do}(X^k := \bar{f}^k(\bar{\mathrm{PA}}^k, \bar{N}^k))$ (Pearl, 2009). This new mechanism determines a conditional that we denote by $p^*(x^k \mid x^{\bar{\mathrm{PA}}^k})$. The interventional distribution is the induced distribution with the new structural assignment, and we denote this by $P := Q^{\mathrm{do}(X^k := \bar{f}(\bar{\mathrm{PA}}^k, \bar{N}^k))}$. If P admits the density p, only the conditional density of X^k changes (e.g., Haavelmo, 1944; Aldrich, 1989; Pearl, 2009; Peters et al., 2017), that is, for $j \neq k$, we have $p(x^j \mid x^{\mathrm{PA}^j}) = q(x^j \mid x^{\mathrm{PA}^j})$, for all x^j and x^{PA^j}. Assume that for the true but unknown distribution Q^* we know the conditional $q^*(x^k \mid x^{\mathrm{PA}^k})$ (e.g., because this was part of the design when generating the data). Due to the factorization property, the intervention $\mathrm{do}(X^k := \bar{f}^k(\bar{\mathrm{PA}}^k, \bar{N}^k))$ can then be represented as a map τ that acts on the density q:

$$\tau(q)(x^1, \ldots, x^d) := \frac{p^*(x^k \mid x^{\bar{\mathrm{PA}}^k})}{q^*(x^k \mid x^{\mathrm{PA}^k})} \cdot q(x^1, \ldots, x^d).$$

Defining $r(x^k) := p^*(x^k \mid x^{\bar{\mathrm{PA}}^k})/q^*(x^k \mid x^{\mathrm{PA}^k})$, this takes the form of (5). As the conditional $p^*(x^k \mid x^{\bar{\mathrm{PA}}^k})$ is fully specified by the intervention, we therefore know the function r. Therefore, the proposed framework allows us to test statements about the distribution $Q^{\mathrm{do}(X^k := \bar{f}(\bar{\mathrm{PA}}^k, \bar{N}^k))}$. We obtain similar expressions when intervening on several variables at the same time.

Similar distributional shifts can be obtained, of course, if the factorization is non-causal (see also Section 3.3), so while our framework contains intervention distributions as a special case, it equally well applies to non-causal models.

B Algorithm for hypothesis testing with unknown distributional shift

This section contains Algorithm 2, which contains details of the hypothesis test when the distributional shift r_q is not known, but can be estimated by an estimator $\hat{r}$. Algorithm 2 is similar to Algorithm 1 but one additionally splits the sample $\mathbf{X}_n$ into two disjoint samples $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ and uses $\mathbf{X}_{n_1}$ for estimating the weights $\hat{r}_{n_1}$, which are then, together with $\mathbf{X}_{n_2}$, 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 2 Testing a target hypothesis with unknown distributional shift and resampling

Input: Data $\mathbf{X}_n$, target sample size m, hypothesis test φ_m, estimator $\hat{r}$ for r_q, and a.

1: Let n_1, n_2 be s.t. $n_1 + n_2 = n$ and $n_1^a = \sqrt{n_2}$
2: $\mathbf{X}_{n_1} \leftarrow X_1, \ldots, X_{n_1}$
3: $\mathbf{X}_{n_2} \leftarrow X_{n_1+1}, \ldots, X_{n_1+n_2}$
4: $\hat{r}_{n_1} \leftarrow$ estimate of r_q based on on $\mathbf{X}_{n_1}$
5: $(i_1, \ldots, i_m) \leftarrow$ sample from $\{1, \ldots, n_2\}^m$ with weights (10) based on $\hat{r}_{n_1}$.
6: $\Psi^{\hat{r}_{n_1},m}_{\texttt{DRPL}}(\mathbf{X}_{n_2}, U) \leftarrow (X_{n_1+i_1}, \ldots, X_{n_1+i_m})$

return $\psi^{\hat{r}}_n(\mathbf{X}_n, U) := \varphi_m(\Psi^{\hat{r}_{n_1},m}_{\texttt{DRPL}}(\mathbf{X}_{n_2}, U))$

C Sampling from $\Psi_{\texttt{DRPL}}$

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

The procedures $\Psi_{\texttt{DRPL}}, \Psi_{\texttt{REPL}}$ and $\Psi_{\texttt{NO-REPL}}$ sample a sequence $(i_1, \ldots, i_m)$ with weight $w_{(i_1,\ldots,i_m)}$

that is given by:

$$
\begin{aligned}
w^{\texttt{DRPL}}_{(i_1,\ldots,i_m)} &= \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})} \quad \text{for distinct } (i_1,\ldots,i_m)\\
w^{\texttt{REPL}}_{(i_1,\ldots,i_m)} &= \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} \quad \text{for all } (i_1,\ldots,i_m)\\
w^{\texttt{NO-REPL}}_{(i_1,\ldots,i_m)} &= \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{j_1=1}^n r(X_{j_1}) \sum\limits_{\substack{j_2=1\\ j_2\neq i_1}}^n r(X_{j_2}) \cdots \sum\limits_{\substack{j_m=1\\ j_m\notin\{i_1,\ldots,i_{m-1}\}}}^n r(X_{j_m})} \quad \text{for distinct } (i_1,\ldots,i_m)
\end{aligned}
$$

The comment 'for distinct $(i_1,\ldots,i_m)$' implies that the weights are zero otherwise. Most statistical software have standard implementations for sampling from $\Psi_{\texttt{REPL}}$ and $\Psi_{\texttt{NO-REPL}}$ (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 $\Psi_{\texttt{DRPL}}$. The first two sampling methods are precise, the third sampling method is approximate.

C.1 Acceptance-rejection sampling with $\Psi_{\texttt{REPL}}$ as proposal

Given a sample $\mathbf{X}_n$, one can sample from $\Psi_{\texttt{DRPL}}$ by acceptance-rejection sampling from $\Psi_{\texttt{REPL}}$, by drawing sequences $(i_1,\ldots,i_m)$ from $\Psi_{\texttt{REPL}}$ until one gets a draw that is distinct, which is then used as the draw from $\Psi_{\texttt{DRPL}}$. This is a valid sampling method for $\Psi_{\texttt{DRPL}}$, because for any distinct sequence $(i_1,\ldots,i_m)$, the conditional probability of drawing $(i_1,\ldots,i_m)$ from $\Psi_{\texttt{REPL}}$ given that $\Psi_{\texttt{REPL}}$ draws a distinct sample equals the probability of $(i_1,\ldots,i_m)$ under $\Psi_{\texttt{DRPL}}$,

$$
\begin{aligned}
\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\ldots,i_m) \mid \Psi_{\texttt{REPL}}\,\text{distinct}, \mathbf{X}_n) &= \frac{\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\ldots,i_m), \Psi_{\texttt{REPL}}\,\text{distinct} \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{REPL}}\,\text{distinct} \mid \mathbf{X}_n)}\\
&= \frac{\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\ldots,i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{REPL}}\,\text{distinct} \mid \mathbf{X}_n)}\\
&= \frac{w^{\texttt{REPL}}_{(i_1,\ldots,i_m)}}{\sum\limits_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} w^{\texttt{REPL}}_{(j_1,\ldots,j_m)}}\\
&= \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum\limits_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} \frac{\sum\limits_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})}{\sum\limits_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}\\
&= w^{\texttt{DRPL}}_{(i_1,\ldots,i_m)}\\
&= \mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1,\ldots,i_m) \mid \mathbf{X}_n)
\end{aligned}
$$

By integrating over $\mathbf{X}_n$, it also holds that $\mathbb{P}_Q(\Psi_{\texttt{REPL}} = (i_1,\ldots,i_m) \mid \Psi_{\texttt{REPL}}\,\text{distinct}) = \mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1,\ldots,i_m))$. One can guarantee that an acceptable sample is drawn quickly. Consider the following strengthening of (A3).

(A4) For all $Q \in \tau^{-1}(H_0)$, there exists $L_Q \in \mathbb{R}$ such that for all $v \geq 1$, $\mathbb{E}_Q[r(X_i)^{v+1}] \leq L_Q^v$. This is trivially satisfied if $r(X_i)$ is Q-a.s. bounded by a constant L.

The following proposition shows that under (A2) and (A4), this acceptance-rejection sampler will quickly accept a sample, in fact the probability of accepting already the first sample from $\Psi_{\texttt{REPL}}$ converges to 1.

Proposition 1. *Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ with densities p and q with respect to a dominating measure μ. Let $r : \mathcal{X} \to (0,\infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) \propto r(x)q(x)$. Let $\mathbf{X}_n$ be a sample* *from Q and let $\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n)$ be the weighted resampling with replacement defined in Section 4.1. Under (A2) and (A4), it holds that*

$$\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \textit{ distinct}) \to 1.$$

In fact, this result also allow us to prove, as a corollary to Theorem 1, pointwise asymptotic level of a test when $\Psi_{\texttt{REPL}}$ is used instead of $\Psi_{\texttt{DRPL}}$.

Corollary 2. *Consider a null hypothesis $H_0 \subseteq \mathcal{P}$ in the target domain. 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 a test φ_m that satisfies (A1) for an $\alpha \in (0, 1)$ and assume that (A2) and (A4) hold. Let $\psi^r_n(\mathbf{X}_n, U) := \varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U))$ denote the output of Algorithm 1, but where $\Psi_{\texttt{REPL}}$ is used instead of $\Psi_{\texttt{DRPL}}$. Then,*

$$\sup_{Q\in\tau^{-1}(H_0)} \limsup_{n\to\infty} \mathbb{P}_Q(\psi^r_n(\mathbf{X}_n, U) = 1) \le \alpha,$$

i.e., ψ^r_n satisfies pointwise asymptotic level α for the hypothesis $\tau(Q^*) \in H_0$.

Proof. We have

$$\begin{aligned}
&\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)) = 1)\\
&\quad= \mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)) = 1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ distinct})\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ distinct})\\
&\qquad+ \mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)) = 1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ not distinct})\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ not distinct}).
\end{aligned}$$

By Proposition 1, this converges to the same limit as $\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)) = 1 \mid \Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ distinct})$, and, as we argue above, this equals $\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1)$. The result then follows from Theorem 1. ☐

C.2 Acceptance-rejection sampling with $\Psi_{\texttt{NO-REPL}}$ as proposal

If m is large compared to n, it may be difficult to sample a distinct sample from using $\Psi_{\texttt{REPL}}$, and so the acceptance rejection scheme in Appendix C.1 make take many attempts to produce a distinct sample. To improve on this, one can use $\Psi_{\texttt{NO-REPL}}$ as a proposal distribution for an acceptance-rejection sampler, which is typically faster, since $\Psi_{\texttt{NO-REPL}}$ has the same support as $\Psi_{\texttt{DRPL}}$. Given a sample $\mathbf{X}_n$, we need to identify M such that

$$\forall (i_1, \ldots, i_m) : \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)} \le M.$$

Let $p_{i_\ell} := \frac{r(X_{i_\ell})}{\sum_j r(X_j)}$. We have

$$\frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)}{\mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (i_1, \ldots, i_m) \mid \mathbf{X}_n)} = \frac{w^{\texttt{DRPL}}_{(i_1,\ldots,i_m)}}{w^{\texttt{NO-REPL}}_{(i_1,\ldots,i_m)}} = \frac{\sum\limits_{j_1} p_{j_1} \sum\limits_{j_2\neq i_1} p_{j_2} \cdots \sum\limits_{j_m\neq i_1,\ldots i_{m-1}} p_{j_m}}{\sum\limits_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}.$$

The denominator does not depend on $i_1, \ldots, i_m$, so an upper bound for the fraction can be obtained by taking $i_1 = \arg\min_i p_i$, and i_2 to be the second smallest etc. Then, the numerator is

$$(1 - 0)(1 - p_{(1)})(1 - p_{(1)} - p_{(2)}) \cdots (1 - p_{(1)} - \ldots - p_{(m-1)}),$$

where $p_{(1)} = \min\{p_1, \ldots, p_n\}$ is the smallest of the weights, $p_{(2)}$ is the second smallest, etc., and we can choose

$$M := \frac{(1 - 0)(1 - p_{(1)})(1 - p_{(1)} - p_{(2)}) \cdots (1 - p_{(1)} - \ldots - p_{(m-1)})}{\sum\limits_{\substack{(j_1,\ldots,j_m)\\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})}.$$

We now proceed with an ordinary acceptance-rejection sampling scheme: We sample a sequence $(\mathbf{i_1}, \ldots, \mathbf{i_m})$ from $\Psi_{\texttt{NO-REPL}}$ and an independent, uniform variable V on the interval $(0, 1)$. We accept $(\mathbf{i_1}, \ldots, \mathbf{i_m})$ if

$$\begin{aligned} V &\leq \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (\mathbf{i_1}, \ldots, \mathbf{i_m}) \mid \mathbf{X_n})}{M \cdot \mathbb{P}_Q(\Psi_{\texttt{NO-REPL}} = (\mathbf{i_1}, \ldots, \mathbf{i_m}) \mid \mathbf{X_n})} \\ &= \frac{(1-0)(1-p_{\mathbf{i_1}})(1-p_{\mathbf{i_1}}-p_{\mathbf{i_2}})\cdots(1-p_{\mathbf{i_1}}-\ldots-p_{\mathbf{i_{m-1}}})}{(1-0)(1-p_{(1)})(1-p_{(1)}-p_{(2)})\cdots(1-p_{(1)}-\ldots-p_{(m-1)})}. \end{aligned}$$

Here we use that the denominator of M cancels with the normalization constant of $\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (\mathbf{i_1}, \ldots, \mathbf{i_m}) \mid \mathbf{X_n})$. If the sample is not accepted, we start over and draw a new sample from $\Psi_{\texttt{NO-REPL}}$.

C.3 Approximate Gibbs sampling starting from $\Psi_{\texttt{NO-REPL}}$

There are cases, where the sampling schemes presented in Appendices C.1 and C.2 are not feasible (this typically is due to m being too large compared to n, which would asymptotically correspond to a violation of (A2)). In that case, one can get an approximate sample of $\Psi_{\texttt{DRPL}}$ by sampling $\Psi_{\texttt{NO-REPL}}$ and shifting it towards $\Psi_{\texttt{DRPL}}$ using a Gibbs sampler (Geman and Geman, 1984).

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

$$\begin{aligned} &\mathbb{P}_Q(\Psi^\ell_{\texttt{DRPL}} = j | \Psi^{-\ell}_{\texttt{DRPL}} = i_{-\ell}) \\ &\qquad = \frac{\mathbb{P}_Q(\Psi_{\texttt{DRPL}} = (i_1, \ldots, j, \ldots, i_m))}{\mathbb{P}_Q(\Psi^{-\ell}_{\texttt{DRPL}} = i_{-\ell})} \\ &\qquad = \frac{r(X_{i_1})\cdots r(X_j)\cdots r(X_{i_m})}{\sum_{v \notin i_{-\ell}} r(X_{i_1})\cdots r(X_v)\cdots r(X_{i_m})} = \frac{r(X_j)}{\sum_{v \notin i_{-\ell}} r(X_v)}, \end{aligned}$$

i.e., the conditional distribution of one index i_ℓ given $i_{-\ell}$ is just a weighted draw among $\{i_1, \ldots, i_m\} \backslash i_{-\ell}$. 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_ℓ by a sample from the conditional given $i_{-\ell}$. Iterating this a large number of times produces an approximate sample from $\Psi_{\texttt{DRPL}}$.

D Proofs

D.1 Proof of Theorem 1

Proof of Theorem 1. Fix any $P \in H_0$ and let $Q \in \tau^{-1}(P) := \tau^{-1}(\{P\})$ and let p and q denote their respective densities with respect to the dominating measure μ. Let $\alpha_\varphi := \limsup_{m\to\infty} \mathbb{P}_P(\varphi_m(\mathbf{Z}_m) = 1)$ be the asymptotic level for a sample $\mathbf{Z}_m$ of size m from P. Then, since by (A1) it holds that $\alpha_\varphi \leq \alpha$, it suffices to prove that $\limsup_{n\to\infty} \mathbb{P}_Q(\psi^r_n(\mathbf{X}_n, U) = 1)) = \alpha_\varphi$.

By assumption $p = \tau(q)$, so $p(x) \propto r(x)q(x)$. Let $\bar{r}$ be the normalized 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 $\Psi_{\texttt{DRPL}}$, defined by (10), samples from the space of distinct sequences $(i_1, \ldots, i_m)$, where every sequence has weight $w_{(i_1,\ldots,i_m)} \propto \prod_{\ell=1}^m r(X_{i_\ell})$. The normalization constant here is the sum over the weights in the entire space of distinct sequences, that is

$$w_{(i_1,\ldots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{\substack{(j_1,\ldots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m r(X_{j_\ell})} = \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{\substack{(j_1,\ldots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})}.$$

Thus, taking an expectation involving $\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$, amounts to evaluating φ_m in all distinct sequences $X_{i_1}, \ldots, X_{i_m}$ and weighting with $w_{(i_1,\ldots,i_m)}$.

$$\mathbb{P}_Q(\varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U)) = 1) = \mathbb{E}_Q \left[\frac{\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell}) \right) \mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}}}{\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(j_1,\ldots,j_m) \\ \text{distinct}}} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})} \right], \quad (12)$$

where we divide with the number of distinct sequences $\frac{n!}{(n-m)!}$ in both numerator and denominator. Applying Lemma 1 to the numerator and denominator, it follows from Slutsky's lemma that the term inside the expectation converges in probability to α_φ. Since the fraction is bounded below and above by 0 and 1 respectively, the convergence in probability implies convergence of the mean,[4] which concludes the proof of Theorem 1. $\square$

Lemma 1 (Distinct draws)**.** *Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ with densities p and q with respect to a dominating measure μ. Let $\bar{r} : \mathcal{X} \to (0, \infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) = \bar{r}(x)q(x)$ and let $(\varphi_m)_m$ be a sequence of tests satisfying that*

$$\lim_{n\to\infty} \mathbb{P}_P[\varphi_m(X_1, \ldots, X_m) = 1] = \alpha_\varphi \in (0, 1).$$

Then, under (A2) and (A3) it holds that

$$\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell}) \right) \mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}} \xrightarrow{Q} \alpha_\varphi \quad \text{as } n \to \infty, \quad (13)$$

and

$$\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell}) \right) \xrightarrow{Q} 1 \quad \text{as } n \to \infty. \quad (14)$$

Proof. To prove (13) and (14), we will proceed in two steps: (A) we first prove that the means converge to the desired limit and (B) we show that the variances in both cases converges to zero. The convergence in probability then follows by Chebyshevs inequality.

Part A (means): Define $\delta_m := \mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}}$ (for the case (13)) or $\delta_m := 1$ (for the

[4] If $(Z_n)_{n=1}^\infty$ is a sequence of random variables such that $Z_n \in [0, 1]$ a.s. and $Z_n \xrightarrow{\mathbb{P}} 0$, we have for all $\varepsilon > 0$ that $\mathbb{E}Z_n \leq \varepsilon P(Z_n \leq \varepsilon) + 1P(Z_n > \varepsilon)$, and since $P(Z_n > \varepsilon) \to 0$, it follows that $\mathbb{E}Z_n \to 0$.

case (14)). Then, in both cases it holds that

$$\begin{aligned}
&\mathbb{E}_Q\left[\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\delta_m\right]\\
&=\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\mathbb{E}_Q\left[\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\delta_m\right]\\
&=\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\int\left(\prod_{\ell=1}^m \bar{r}(x_{i_\ell})q(x_{i_\ell})\right)\delta_m \mathrm{d}\mu^m(x_{i_1},\ldots,x_{i_m})\\
&=\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\int\left(\prod_{\ell=1}^m p(x_{i_\ell})\right)\delta_m \mathrm{d}\mu^m(x_{i_1},\ldots,x_{i_m})\\
&=\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\mathbb{E}_P[\delta_m]\\
&=\mathbb{E}_P[\delta_m]
\end{aligned}$$

In the second last equality, we used 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 $\frac{n!}{(n-m)!}$. Conclusively, the expectation either converges to α_φ in the case (13) or to 1 in the case (14).

Part B (variances): We begin by expressing $\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)$ as a U-statistic (Serfling, 1980). A U-statistic has the form

$$\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} h(Z_{i_1},\ldots,Z_{i_m}) \tag{15}$$

for some symmetric function $h(z_1,\ldots,z_m)$ (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

$$\mathbb{V}_Q\left(\frac{1}{\frac{n!}{(n-m)!}}\sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}}\left(\prod_{\ell=1}^m \bar{r}(X_{i_\ell})\right)\right)=\binom{n}{m}^{-1}\sum_{v=1}^m\binom{m}{v}\binom{n-m}{m-v}\zeta_v \tag{16}$$

where

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

We now bound ζ_v from above by the second moment as follows

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

Moreover, using that δ_m is upper bounded by 1, we get for both cases (14) and (13) that

$$\begin{aligned}
\zeta_v &\le \mathbb{E}_Q\left[\mathbb{E}_Q[h_m(X_{i_1},\ldots,X_{i_m}) \mid X_{i_1},\ldots,X_{i_v}]^2\right]\\
&\le \mathbb{E}_Q\left[\mathbb{E}_Q\left[\prod_{\ell=1}^m \bar{r}(X_{i_\ell}) \mid X_{i_1},\ldots,X_{i_v}\right]^2\right].
\end{aligned} \tag{17}$$

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

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

where the last equality follows because

$$
\mathbb{E}_Q[\bar{r}(X_{i_\ell})] = \int \bar{r}(x) q(x) \mathrm{d}\mu(x) = \int p(x) \mathrm{d}\mu(x) = 1.
$$

Next, starting from the bound in (17) we get that

$$
\begin{aligned}
\zeta_v &\leq \mathbb{E}_Q\left[\left(\prod_{\ell=1}^{v} \bar{r}(X_{i_\ell})\right)^2\right] \\
&= \prod_{\ell=1}^{v} \mathbb{E}_Q\left[\bar{r}(X_{i_\ell})^2\right] \\
&= \mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right]^v.
\end{aligned}
$$

Here, we again used the independence of the distinct terms and the fact that $\mathbb{E}_Q\left[\bar{r}(X_{i_\ell})\right] = 1$. Plugging this into (16), we get

$$
\begin{aligned}
&\mathbb{V}_Q\left(\frac{1}{\frac{n!}{(n-m)!}} \sum_{\substack{(i_1,\ldots,i_m) \\ \text{distinct}}} \left(\prod_{\ell=1}^{m} \bar{r}(X_{i_\ell})\right)\right) \\
&\leq \binom{n}{m}^{-1} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} \mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right]^\ell.
\end{aligned}
$$

By (A3), $\mathbb{E}_Q\left[\bar{r}(X_{i_1})^2\right] < \infty$, so Lemma 2 implies that this converges to 0 for $n \to \infty$. This shows that the variance converges to zero in both cases (14) and (13), which completes the proof of Lemma 1. □

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

$$
\lim_{n \to \infty} \frac{1}{\binom{n}{m}} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} K^\ell = 0. \tag{18}
$$

Remark. *The Chu-Vandermonde identity states that $\frac{1}{\binom{n}{m}} \sum_{\ell=0}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell} = 1$. In light of this identity, one may be surprised that when including the exponentially growing term, K^ℓ, the sum vanishes. The reason is that the summation in (18) 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, m approaches ∞.*

Proof. Denote by s_ℓ the ℓ'th summand, i.e.,

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

It then holds for all $\ell \in \{1, \dots, m-1\}$ that

$$\begin{aligned}
\frac{s_{\ell+1}}{s_\ell} &= \frac{\binom{m}{\ell+1}\binom{n-m}{m-\ell-1}K^{\ell+1}}{\binom{m}{\ell}\binom{n-m}{m-\ell}K^\ell} \\
&= \frac{\frac{m!}{(\ell+1)!(m-\ell-1)!}\frac{(n-m)!}{(m-\ell-1)!(n-2m+\ell+1)!}}{\frac{m!}{\ell!(m-\ell)!}\frac{(n-m)!}{(m-\ell)!(n-2m+\ell)!}}K \\
&= \frac{(m-\ell)^2}{\ell(n-2m+\ell+1)}K \\
&\leq \frac{m^2}{(n-2m+2)}K
\end{aligned}$$

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, m sufficiently large it holds for all $\ell \in \{1, \dots, m-1\}$ that $\frac{s_{\ell+1}}{s_\ell} \leq c$. 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 (18) as follows

$$\begin{aligned}
\frac{1}{\binom{n}{m}} \sum_{\ell=1}^{m} \binom{m}{\ell}\binom{n-m}{m-\ell}K^\ell &= \frac{1}{\binom{n}{m}} \sum_{\ell=1}^{m} s_\ell \\
&< \frac{1}{1-c}\frac{s_1}{\binom{n}{m}} \\
&= \frac{K}{1-c}\frac{m\binom{n-m}{m-1}}{\binom{n}{m}} \\
&= \frac{K}{1-c}\frac{m^2}{n}\frac{\binom{n-m}{m-1}}{\binom{n-1}{m-1}},
\end{aligned} \tag{19}$$

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 $\frac{\binom{n-m}{m-1}}{\binom{n-1}{m-1}} \leq 1$, it immediately follows that (19) converges to zero. This completes the proof of Lemma 2. $\square$

Lemma 3. *Define for all $n, m \in \mathbb{N}$ the function*

$$g(n,m) \coloneqq \frac{n!}{(n-m)!} n^{-m}.$$

Then, it holds that

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

Proof. First, apply the Stirling approximation to get for m, n sufficiently large that

$$\begin{aligned}
g(n,m) &\sim n^{n+\frac{1}{2}} \cdot e^{-n} \cdot (n-m)^{m-n-\frac{1}{2}} \cdot e^{n-m} \cdot n^{-m} \\
&= n^{n-m+\frac{1}{2}} \cdot (n-m)^{m-n-\frac{1}{2}} \cdot e^{-m} \\
&= \exp\{(n-m+\tfrac{1}{2})\log(n) + (m-n-\tfrac{1}{2})\log(n-m) - m\}.
\end{aligned}$$

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

$$\begin{aligned}
g(n, n^q) &\sim \exp\{(n - n^q + \tfrac{1}{2})\log(n) + (n^q - n - \tfrac{1}{2})\log(n - n^q) - n^q\} \\
&= \exp\{(n - n^q + \tfrac{1}{2})\log(n) + (n^q - n - \tfrac{1}{2})[\log(n) + \log(1 - n^{q-1})] - n^q\} \\
&= \exp\{(n^q - n - \tfrac{1}{2})\log(1 - n^{q-1}) - n^q\}.
\end{aligned}$$

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

$$\log(1 + n^{q-1}) = -n^{q-1} - \tfrac{1}{2}n^{2(q-1)} + O(n^{3(q-1)}),$$

which results in

$$\begin{aligned}
g(n, n^q) &\sim \exp\{(n^q - n - \tfrac{1}{2})\log(1 - n^{q-1}) - n^q\} \\
&= \exp\{(n^q - n - \tfrac{1}{2})(-n^{q-1} - \tfrac{1}{2}n^{2(q-1)} + O(n^{3(q-1)})) - n^q\} \\
&= \exp\{-n^{2q-1} - \tfrac{1}{2}n^{3q-2} + n^q + \tfrac{1}{2}n^{3q-2} + \tfrac{1}{2}n^{q-1} + \tfrac{1}{4}n^{2q-2} + O(n^{2q-1}) - n^q\} \\
&= \exp\{-\tfrac{1}{2}n^{2q-1} + O(n^{3q-2})\}.
\end{aligned}$$

From this we see that

$$\lim_{n\to\infty} g(n, n^q) = \begin{cases} 0 & \text{if } q \in (\tfrac{1}{2}, \infty) \\ \exp(-\tfrac{1}{2}) & \text{if } q = \tfrac{1}{2} \\ 1 & \text{if } q \in [0, \tfrac{1}{2}). \end{cases}$$

This completes the proof of Lemma 3. □

D.2 Proof of Theorem 2

Proof. The proof will follow the arguments in the proof of Theorem 1 (for sampling without replacement) but 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 $\hat{r}_{n_1}$.

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 μ.

We begin by recalling the details for the sample splitting procedure described in Algorithm 2: $\mathbf{X}_n$ is split into two disjoint data sets $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ of sizes n_1, n_2, where $n_1 + n_2 = n$. In particular, we assume that $n_1^a = \sqrt{n_2}$, since this ensures that by assumption $m = min(n^a, \sqrt{n})$, we also have $m = o(n_1^a)$ and $m = o(\sqrt{n_2})$.[5] Moreover, we use $\mathbf{X}_{n_1}$ to fit an estimator $\hat{r}_{n_1}$ of r_q and then use $\mathbf{X}_{n_2}$ for the resampling. When taking expectations over $\mathbf{X}_{b_1}$ we use the notation $\mathbb{E}_{Q_{n_1}}$. Similarly, $\mathbb{E}_{Q_{n_2}}$ denotes an expectation over $\mathbf{X}_{n_2}$. We write $\mathbb{E}_Q$ when taking expectations with respect to the entire sample $\mathbf{X}_n$.

Using the same argument as we used to derive (12) in the proof of Theorem 1, we get that

$$\mathbb{P}_Q(\varphi_m(\Psi^{\hat{r}_{n_1}}(\mathbf{X}_{n_2}, U) = 1)) = \mathbb{E}_Q \left[\frac{\frac{1}{n_2^m} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \left(\prod\limits_{\ell=1}^m \hat{r}_{n_1}(X_{i_\ell}) \right) \mathbb{1}_{\{\varphi_m(X_{i_1},\dots,X_{i_m})=1\}}}{\frac{1}{n_2^m} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \prod\limits_{\ell=1}^m \hat{r}_{n_1}(X_{i_\ell})} \right], \tag{20}$$

where X_{i_ℓ} are samples from $\mathbf{X}_{n_2}$. As in the proof of Lemma 1, we prove the convergence in probability of the nominator and denominator in (20) separately. Again, we do this in two steps: (A) show that the means convergence to the desired quantity and (B) show that the variances converge to zero.

[5] When $n_1^a = \sqrt{n_2}$ and $n_1 + n_2 = n$, we have $n = n_1^{2a} + n_1$ and $n = n_2 + n_2^{1/(2a)}$. If $a > 1/2$ we have $m = o(\sqrt{n}) = o(\sqrt{n_1^{2a}}) = o(n_1^a)$, and $m = o(\sqrt{n}) = o(\sqrt{n_2})$. Similar arguments apply if $a < 1/2$.

Depending on whether we consider the nominator or denominator case, we define either $\delta_m := \mathbb{1}_{\{\varphi_m(X_{i_1},\ldots,X_{i_m})=1\}}$ or $\delta_m = 1$. Furthermore, to ease notation we introduce for any function $r : \mathcal{X} \to (0,\infty)$ the following short-hand

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

Part A (means): Using the independence between $\mathbf{X}_{n_1}$ and $\mathbf{X}_{n_2}$ we get that

$$\begin{aligned}
\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] &= \frac{1}{n_2^m} \sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \mathbb{E}_Q\left[\left(\prod_{\ell=1}^m r_q(X_{i_\ell}) \right) \left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right) \delta_m \right] \\
&= \frac{1}{n_2^m} \sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \mathbb{E}_{Q_2}\left[\left(\prod_{\ell=1}^m r_q(X_{i_\ell}) \right) \mathbb{E}_{Q_1}\left[\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right] \delta_m \right]. \qquad (21)
\end{aligned}$$

We emphasize that the expectation $\mathbb{E}_{Q_1}$ only averages over the randomness in estimating $\hat{r}$, and does take expectations over X_{i_l}, since this is a draw from Q_{n_2}. Furthermore, using the intermediate result (22) we get the following upper bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] \leq \mathbb{E}_Q\left[M(r_q)\right](1+\varepsilon(n_1))$$

and lower bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})\right] \geq \mathbb{E}_Q\left[M(r_q)\right](1-\varepsilon(n_1)).$$

Since both n_1 and n_2 (and in particular $m = o(\sqrt{n_2})$) we can apply the convergence shown in Part A in the proof of Lemma 1 to get that the means $\mathbb{E}_Q[M(\hat{r}_{n_1})]$ of the denominator and numerator converge to 1 and α_φ respectively.

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

$$\begin{aligned}
\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] &= \mathbb{E}_Q\left[\frac{1}{n_2^{2m}} \sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \sum_{\substack{i'_1,\ldots,i'_m\\ \text{distinct}}} \left(\prod_{\ell=1}^m \hat{r}_{n_1}(X_{i_\ell}) \right) \left(\prod_{\ell=1}^m \hat{r}_{n_1}(X_{i'_\ell}) \right) \delta_m \right] \\
&= \mathbb{E}_{Q_2}\Bigg[\frac{1}{n_2^{2m}} \sum_{\substack{(i_1,\ldots,i_m)\\ \text{distinct}}} \sum_{\substack{i'_1,\ldots,i'_m\\ \text{distinct}}} \left(\prod_{\ell=1}^m r(X_{i_\ell}) \right) \left(\prod_{\ell=1}^m r(X_{i'_\ell}) \right) \\
&\qquad \mathbb{E}_{Q_1}\left[\left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right) \left(\prod_{\ell=1}^m \frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})} \right) \right] \delta_m \Bigg].
\end{aligned}$$

Using the intermediate result (23) we get the following upper bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] \leq \mathbb{E}_Q\left[M(r_q)^2\right](1+\varepsilon(n_1))^2$$

and lower bound

$$\mathbb{E}_Q\left[M(\hat{r}_{n_1})^2\right] \geq \mathbb{E}_Q\left[M(r_q)^2\right](1-\varepsilon(n_1))^2.$$

In Part A and Part B of the proof of Lemma 1 we have shown that $\lim_{n\to\infty} \mathbb{V}_Q(M(r_q)) = 0$ and $\lim_{n\to\infty} \mathbb{E}_Q(M(r_q)) = 1$ or α_φ (for denominator and nominator respectively). Hence, combining this with the above bounds shows that also $\lim_{n\to\infty} \mathbb{V}_Q(M(\hat{r}_{n_1})) = 0$.

Intermediate results: Let $\varepsilon(n_1) := \sup_{x\in\mathcal{X}} \mathbb{E}_{Q_{n_1}} \left| \left(\frac{\hat{r}_{n_1}(x)}{r_q(x)}\right)^{n_1^a} - 1\right|$. Then, for n_1 sufficiently large and using Jensen's inequality, it holds Q_{n_2}-a.s. that

$$
\begin{aligned}
\left| \mathbb{E}_{Q_{n_1}} \left[\prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right] - 1 \right| &\leq \mathbb{E}_{Q_{n_1}} \left[\left| \prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} - 1 \right| \right] \\
&\leq \sup_{x\in\mathcal{X}} \mathbb{E}_{Q_{n_1}} \left[\left| \left(\frac{\hat{r}_{n_1}(x)}{r_q(x)} \right)^{m} - 1 \right| \right] \\
&\leq \sup_{x\in\mathcal{X}} \mathbb{E}_{Q} \left[\left| \left(\frac{\hat{r}_{n_1}(x)}{r_q(x)} \right)^{n_1^a} - 1 \right| \right] \\
&= \varepsilon(n_1),
\end{aligned}
\tag{22}
$$

where we used that by assumption, $m = o(n_1^a)$ when $n_1 \to \infty$, so for n_1 sufficiently large, $n_1^a > m$.[6] Similarly, we get Q_{n_2}-a.s. that

$$
\begin{aligned}
&\left| \mathbb{E}_{Q_{n_1}} \left[\left(\prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right) \left(\prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})} \right) \right] - 1 \right| \\
&\leq \mathbb{E}_{Q_{n_1}} \left[\left| \left(\prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i_\ell})}{r_q(X_{i_\ell})} \right) \left(\prod_{\ell=1}^{m} \frac{\hat{r}_{n_1}(X_{i'_\ell})}{r_q(X_{i'_\ell})} \right) - 1 \right| \right] \\
&\leq \sup_{x\in\mathcal{X}} \mathbb{E}_{Q_{n_1}} \left[\left| \left(\frac{\hat{r}_{n_1}(x)}{r_q(x)} \right)^{2m} - 1 \right| \right] \\
&\leq \sup_{x\in\mathcal{X}} \mathbb{E}_{Q} \left[\left| \left(\frac{\hat{r}_{n_1}(x)}{r_q(x)} \right)^{n_1^a} - 1 \right| \right] \\
&= \varepsilon(n_1),
\end{aligned}
\tag{23}
$$

using that for n_1 sufficiently large, $n_1^a > 2m$. This completes the proof of Theorem 2. □

D.3 Proof of Proposition 1

Proof. When sampling with replacement, $\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)$ contains non-distinct draws with positive probability (assuming, wlog, that m is not 1). Yet, we show that $\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U)$ distinct) approaches 1 as $m \to \infty$. By assumption $p = \tau(q)$, so $p(x) \propto r(x)q(x)$. Let $\bar{r}$ be the normalized version of r satisfying $p(x) = \bar{r}(x)q(x)$. The probability of drawing a sequence $X_{i_1}, \ldots, X_{i_m}$, $w_{(i_1,\ldots,i_m)}$, is defined by (10) as the product of weights r factors:

$$
w_{(i_1,\ldots,i_m)} = \frac{\prod_{\ell=1}^m r(X_{i_\ell})}{\sum_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m r(X_{j_\ell})} = \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{(j_1,\ldots,j_m)} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})},
$$

where the sum over $(j_1, \ldots, j_m)$ in the denominator is over all sequences of length m, distinct or not. The probability of drawing a non-distinct sequence, is the sum of the weights corresponding

[6]This holds because 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 $0 \leq c \leq 1$. Thus, for all $c \geq 0$ it holds that $|c^{m+k} - 1| \geq |c^m - 1|$.

to all non-distinct sequences $w_{(i_1,\dots,i_m)}$. Therefore,

$$
\begin{aligned}
&\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ not distinct}) \\
&= \mathbb{E}_Q\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} w_{(i_1,\dots,i_m)}\right] \\
&= \mathbb{E}_Q\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \frac{\prod_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum_{(j_1,\dots,j_m)} \prod_{\ell=1}^m \bar{r}(X_{j_\ell})}\right] \\
&= \mathbb{E}_Q\left[\frac{\sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \prod\limits_{\ell=1}^m \bar{r}(X_{i_\ell})}{\sum\limits_{(i_1,\dots,i_m)} \prod\limits_{\ell=1}^m \bar{r}(X_{i_\ell})}\right] \\
&= \mathbb{E}_Q\left[\frac{\frac{1}{n^m} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \prod\limits_{\ell=1}^m \bar{r}(X_{i_\ell})}{\frac{1}{n^m} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \prod\limits_{\ell=1}^m \bar{r}(X_{i_\ell}) + \frac{1}{n^m} \sum\limits_{\substack{(i_1,\dots,i_m)\\ \text{distinct}}} \prod\limits_{\ell=1}^m \bar{r}(X_{i_\ell})}\right].
\end{aligned} \tag{24}
$$

Observe that this expectation is taken both over $\mathbf{X}_n$ and U. By Lemma 4, the numerator of (24) (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 (A3), 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 $\mathbb{P}_Q(\Psi^{r,m}_{\texttt{REPL}}(\mathbf{X}_n, U) \text{ not distinct}) \to 0$. $\square$

Lemma 4 (Non-distinct draws). *Let $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ with densities p and q with respect to a dominating measure μ. Let $\bar{r} : \mathcal{X} \to (0, \infty)$ satisfy for all $x \in \mathcal{X}$ that $p(x) = \bar{r}(x)q(x)$. Then, under (A2) and (A4) it holds that*

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

In particular, since the integrand is non-negative, this implies that

$$
\frac{1}{n^m} \sum_{\substack{(i_1,\dots,i_m)\\ \textit{not distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell}) \xrightarrow{L^1} 0 \quad \textit{as } n \to \infty.
$$

Proof. We begin by rewriting the sum in terms of the number of distinct draws k, i.e. there are k distinct elements among $i_1, \dots, i_m$. 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 number of occurences of each index $r_1, \dots, r_k$, i.e. j_ℓ appears r_ℓ times.

$$
\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \prod_{\ell=1}^m \bar{r}(X_{i_\ell}) = \sum_{k=1}^{m-1} \sum_{\substack{j_1,\dots,j_k\\ \text{distinct}}} \sum_{\substack{r_1,\dots,r_k>0\\ r_1+\dots+r_k=m}} \prod_{\ell=1}^k \bar{r}(X_{j_\ell})^{r_\ell}. \tag{25}
$$

Using the independence across distinct samples, this implies that

$$\mathbb{E}_Q\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar r(X_{i_\ell})\right]=\sum_{k=1}^{m-1}\sum_{\substack{j_1,\dots,j_k\\ \text{distinct}}}\sum_{\substack{r_1,\dots,r_k>0\\ r_1+\dots+r_k=m}}\prod_{\ell=1}^{k}\mathbb{E}_Q\left[\bar r(X_{j_\ell})^{r_\ell}\right]. \tag{26}$$

Next, use the uniform bound on the weights given in (A4) and the fact that $\mathbb{E}_Q[\bar r(X_i)^t]=\int \bar r(x_i)q(x_i)\bar r(x_i)^{t-1}\mathrm{d}\mu(x_i)=\int p(x_i)\bar r(x_i)^{t-1}\mathrm{d}\mu(x_i)=\mathbb{E}_P[\bar r(X_i)^{t-1}]$ to get for all $i\in\{1,\dots,n\}$ and all $t\in\{1,\dots,m-1\}$ that

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

Together with (26) this results in

$$\begin{aligned}\mathbb{E}\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar r(X_{i_\ell})\right]&\le\sum_{k=1}^{m-1}\sum_{\substack{j_1,\dots,j_k\\ \text{distinct}}}\sum_{\substack{r_1,\dots,r_k>0\\ r_1+\dots+r_k=m}}L^{m-k}\\&=\sum_{k=1}^{m-1}\binom{n}{k}\pi(m,k)L^{m-k}\\&\le\sum_{k=1}^{m-1}\binom{n}{k}\tilde\pi(m,k)L^{m-k},\end{aligned}\tag{27}$$

where $\pi(m,k)$ is the number of words of length m using k letters such that each is used at least once and

$$\tilde\pi(m,k):=k^{m-k}\frac{m!}{k!}.$$

We have that $\pi(m,k)\le\tilde\pi(m,k)$ because $\pi(m,k)\le k^m=k^{m-k}k^k\le k^{m-k}(k+1)\cdots m=\tilde\pi(m,k)$. Then, with $s_k:=\binom{n}{k}\tilde\pi(m,k)L^{m-k}$ it holds that

$$\frac{s_{k+1}}{s_k}=\frac{\binom{n}{k+1}\tilde\pi(m,k+1)L^{m-k-1}}{\binom{n}{k}\tilde\pi(m,k)L^{m-k}}=\frac{1}{L}\frac{n-k}{k+1}\frac{1}{k+1}\frac{(k+1)^{m-k-1}}{k^{m-k}}\ge\frac{1}{L}\frac{n-m+1}{(m-1)^2}=:c.$$

By (A2) (i.e., $m=o(\sqrt{n})$) it holds for n,m sufficiently large that $c>1$. Iterating this inequality, we get (again for n,m sufficiently large) that $s_k\le c^{-(m-1-k)}s_{m-1}$ which we can plug into (27) to get

$$\mathbb{E}\left[\sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}}\prod_{\ell=1}^{m}\bar r(X_{i_\ell})\right]\le\sum_{k=1}^{m-1}s_k\le s_{m-1}\sum_{k=1}^{m-1}c^{-(m-1-k)}=s_{m-1}\sum_{k=0}^{m-2}c^{-k}\le s_{m-1}\frac{1}{1-\frac{1}{c}}. \tag{28}$$

In the last inequality, we use the trivial bound $\sum_{k=0}^{m-1}c^{-k}<\sum_{k=0}^{\infty}c^{-k}$ which is a geometric series, which converges because $0<c^{-1}<1$. Finally, observe that

$$\begin{aligned}n^{-m}s_{m-1}&=\binom{n}{m-1}\tilde\pi(m,m-1)L\\&\le n^{-m}\frac{n!}{(n-m)!}\frac{1}{(m-1)!}m(m-1)L\\&=\underbrace{g(n,m)}_{:=n^{-m}\frac{n!}{(n-m)!}}\frac{m}{(m-1)!}L,\end{aligned}$$

which together with Lemma 3 converges to zero (by the assumption $m = o(\sqrt{n})$). Therefore, we have that

$$\mathbb{E}\left[n^{-m} \sum_{\substack{(i_1,\dots,i_m)\\ \text{not distinct}}} \prod_{\ell=1}^{m} \bar{r}(X_{i_\ell})\right] \leq n^{-m} s_{m-1} \frac{1}{1-\frac{1}{c}} \to 0,$$

which completes the proof of Lemma 4. $\square$

D.4 Hypothesis in the observation domain

As discussed in Section 2.3, the proposed procedure in Section 4.1 can also be used to construct a test for a hypothesis $H_0^{\mathcal{Q}}$ in the observational domain, satisfying the same theoretical guarantees.

Corollary 3 (Hypothesis testing in the observational domain)**.** *Consider hypotheses $H_0^{\mathcal{Q}} \subseteq \mathcal{Q}$ and $H_0^{\mathcal{P}} \subseteq \mathcal{P}$ in the observational and in the target domain, respectively. Let $\tau : \mathcal{Q} \to \mathcal{P}$ be a distributional shift for which there exist a known map $r : \mathcal{X} \to (0, \infty)$ and a set A satisfying for all $q \in \mathcal{Q}$ and all $x \in \mathcal{Z}$ it holds that $\tau(q)(x) \propto r(x^A)q(x)$, see (5). Assume $\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}}$. Consider a test φ_m for the hypothesis $H_0^{\mathcal{P}}$ and assume that (A1), (A2) and (A3) hold. Let $\psi_n^r(\mathbf{X}_n, U) \coloneqq \varphi_m(\Psi^{r,m}_{\texttt{DRPL}}(\mathbf{X}_n, U))$ denote the output of Algorithm 1. Then*

$$\sup_{Q \in H_0^{\mathcal{Q}}} \limsup_{n\to\infty} \mathbb{P}_Q(\psi_n^r(\mathbf{X}_n, U) = 1) \leq \alpha,$$

i.e., ψ_n^r satisfies pointwise asymptotic level α for the hypothesis $H_0^{\mathcal{Q}}$.

Clearly, the condition $H_0^{\mathcal{Q}} \subseteq \tau(H_0^{\mathcal{P}})$ is satisfied when $H_0^{\mathcal{P}} = \tau(H_0^{\mathcal{Q}})$. This is the case for the conditional independence test described in Section 3.3, for example.

Proof. We have

$$\tau(H_0^{\mathcal{Q}}) \subseteq H_0^{\mathcal{P}} \;\Rightarrow\; H_0^{\mathcal{Q}} \subseteq \tau^{-1}(H_0^{\mathcal{P}})$$

and therefore

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

The statement then follows from Theorem 1. $\square$

E Analyzing (A3) in a linear Gaussian model

In this section, we show conditions for assumption (A3) to be satisfied when we consider the shift that changes a Gaussian conditional into an independent Gaussian target distribution.

Proposition 2. *Consider a linear Gaussian setting where $Y = X + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and $X \sim \mathcal{N}(0, \sigma_X^2)$, with $\sigma_\varepsilon, \sigma_X$ known. Assume that we are interested in the distributional shift that replaces the conditional $q(y|x)$ with an independent $\mathcal{N}(0, \sigma^2)$ 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 (A3) is satisfied if and only if

$$\sigma^2 < 2(\sigma_\varepsilon^2 - \sigma_X^2).$$

Proof. We begin by directly expanding the second moment of the factor r under the observational distribution Q as follows,

$$
\begin{aligned}
\mathbb{E}_Q\left[r(X,Y)^2\right] &= \mathbb{E}_Q\left[\left(\frac{p(Y)}{q(Y\mid X)}\right)^2\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\left(\frac{Y-X}{\sigma_\varepsilon}\right)^2-\left(\frac{Y}{\sigma}\right)^2\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\left(\frac{\varepsilon}{\sigma_\varepsilon}\right)^2-\left(\frac{X+\varepsilon}{\sigma}\right)^2\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(\varepsilon^2\left(\frac{1}{\sigma_\varepsilon^2}-\frac{1}{\sigma^2}\right)-\frac{X^2}{\sigma^2}-\frac{2X\varepsilon}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{X^2}{\sigma^2}-\frac{2X\varepsilon}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\frac{\sigma^2-\sigma_\varepsilon^2}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{X^2}{\sigma^2}-\frac{2\varepsilon}{\sigma^2}X-\frac{\varepsilon^2}{\sigma^2}+\frac{\varepsilon^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\frac{\sigma^2-\sigma_\varepsilon^2}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W+\frac{\varepsilon^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\frac{\sigma^2-\sigma_\varepsilon^2}{\sigma^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right], \qquad (29)
\end{aligned}
$$

where $W \coloneqq (X/\sigma_X + \varepsilon/\sigma_X)^2$. Next, observe that, conditioned on ε, W has a non-central $\chi^2_{(1)}$-distribution with mean ε/σ_X. The moment generating function of W is given by $M_W(t) = (1-2t)^{-1/2}\exp\left(\frac{\varepsilon^2}{\sigma_X^2}\frac{t}{1-2t}\right)$ for all $t < 1/2$. Hence, continuing the computation in (29) and by conditioning on ε, we get

$$
\begin{aligned}
\mathbb{E}_Q\left[r(X,Y)^2\right] &= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[\mathbb{E}_Q\left[\exp\left(-\frac{\sigma_X^2}{\sigma^2}W\right)\Big|\varepsilon\right]\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\mathbb{E}_Q\left[M_W\left(-\frac{\sigma_X^2}{\sigma^2}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}\mathbb{E}_Q\left[\exp\left(\frac{\varepsilon^2}{\sigma_X^2}\frac{-\sigma_X^2}{\sigma^2}\frac{1}{1+2\frac{\sigma_X^2}{\sigma^2}}\right)\exp\left(\frac{\varepsilon^2}{\sigma_\varepsilon^2}\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}\mathbb{E}_Q\left[\exp\left(-\frac{\varepsilon^2}{\sigma_\varepsilon^2}\left(1-\frac{\sigma_\varepsilon^2}{\sigma^2+2\sigma_X^2}\right)\right)\right] \\
&= \frac{\sigma_\varepsilon^2}{\sigma^2}\left(1+2\frac{\sigma_X^2}{\sigma^2}\right)^{-\frac{1}{2}}M_S\left(1-\frac{\sigma_\varepsilon^2}{\sigma^2+2\sigma_X^2}\right),
\end{aligned}
$$

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