Testing Partially-Identifiable Causal Queries Using Ternary Tests

Sourbh Bhadane[1] **Joris M. Mooij**[1] **Philip Boeken**[2] **Onno Zoeter**[3]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]Department of Mathematics, Vrije Universiteit Amsterdam
[3]Booking.com, The Netherlands

Abstract

We consider hypothesis testing of binary causal queries using observational data. Since the mapping of causal models to the observational distribution that they induce is not one-to-one, in general, causal queries are often only partially identifiable. When binary statistical tests are used for testing partially identifiable causal queries, their results do not translate in a straightforward manner to the causal hypothesis testing problem. We propose using ternary (three-outcome) statistical tests to test partially identifiable causal queries. We establish testability requirements that ternary tests must satisfy in terms of uniform consistency and present equivalent topological conditions on the hypotheses. To leverage the existing toolbox of binary tests, we prove that obtaining ternary tests by combining binary tests is complete. Finally, we demonstrate how topological conditions serve as a guide to construct ternary tests for two concrete causal hypothesis testing problems, namely testing the instrumental variable (IV) inequalities and comparing treatment efficacy.

1 INTRODUCTION

A causal query is deemed identifiable from observational data whenever causal models that induce the same observational distribution also agree on the query of interest. For hypothesis testing of binary causal queries, such as testing the existence of an edge in a causal graph or testing whether a causal effect is 0 or not, identification allows us to translate the results of a statistical test based on observational data to implications about causal hypotheses. For identifiable binary causal queries, the set of observational distributions induced by causal models where the binary causal query evaluates to 0, is disjoint from those induced by causal models where the binary causal query evaluates to 1. However, often causal queries of interest are "not identifiable" from observational data because of reasons such as the presence of unobserved confounders, positivity violations, etc. In such cases, the aforementioned sets of induced observational distributions overlap. If the overlap is partial, then the causal query is *partially identifiable*. If the sets completely overlap, the causal query is deemed *non-identifiable* since every distribution could be induced by causal models corresponding to both values of the causal query.

When binary statistical tests are used to test partially identifiable causal queries, their results do not translate in a straightforward manner to the causal hypothesis testing problem. For example, consider a binary test based on observational data such that the statistical null hypothesis is the set of observational distributions induced by causal models where the causal query evaluates to 0. In this case, rejecting the null implies that the causal query, evaluated on the underlying causal model, evaluates to 1. However, not rejecting the null presents an identifiability issue since the null includes the non-empty intersection of the induced sets of observational distributions. Therefore, not rejecting the null only allows concluding "unidentifiable". See Bhadane et al. (2025) for an example of a statistical test for the absence of a direct effect that leads to a similar conclusion.

In this work, we propose testing partially identifiable causal queries using statistical tests that output one of three outcomes: "reject the null", "don't reject the null", and "unidentifiable", where the former two correspond to the set differences of the induced observational distributions and the latter corresponds to the intersection. Clearly, a ternary test provides more fine-grained information than a binary test. In addition, translating the result of a ternary test to the causal hypothesis testing problem is straightforward; if the ternary test outputs either "reject the null" or "don't reject the null", the conclusion for the causal hypotheses is exact, otherwise it is unidentifiable from observational data.

We extend the binary hypothesis testing framework to test-

Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence (UAI 2026), PMLR 337:549–566.

ing overlapping hypotheses with three outcomes which can be viewed as testing three disjoint hypotheses with three outcomes. We consider an error to be when the output of a ternary test is different from the hypothesis that contained the distribution that samples were drawn from. We first define error-guarantee desiderata for ternary tests. In binary hypothesis testing, these are made in terms of asymptotic consistency and error-control (Dembo and Peres, 1994; Ermakov, 2017; Genin and Kelly, 2017; Genin, 2018; Boeken et al., 2026) and then related to topological conditions on the hypotheses in a suitable topology of probability measures. We follow suit by proposing analogous equivalent topological conditions for existence of 'good' ternary tests. Given the vast toolbox of binary tests developed so far, we study ternary tests that are constructed by combining binary tests, which we call a "two-stage ternary test". We show via topological conditions on the hypotheses that two-stage ternary testing is complete in the sense that it attains the same error control as that of ternary tests that control "column-wise" errors. We demonstrate how topological conditions on hypotheses can guide a practitioner in constructing ternary tests by considering two concrete applications to testing causal queries, namely testing the instrumental variable (IV) inequalities (Pearl, 1995) expressed as conditions on the interventional Markov kernel using observational data, and comparing the treatment efficacy of a treatment from observational data against a threshold that might, for example, correspond to the estimated efficacy of a competing treatment obtained from a large randomized controlled trial.

1.1 RELATED WORK

Ternary hypothesis testing has often been proposed in the context of mitigating shortcomings of the Neyman-Pearson null hypothesis significance testing framework. One of the earliest mentions can be found in the works of Neyman (Neyman, 1976), Tukey (Tukey, 1991; Jones and Tukey, 2000) and Hays (1963). In particular, Esteves et al. (2016) show that statistical tests that are logically coherent and optimal are ternary and not binary. Berg (2004) considers ternary tests for simple binary hypothesis testing and provides an optimal likelihood ratio test for the same. Izbicki et al. (2023) considers a region-based test that constructs confidence regions based on data and defines a ternary test based on whether the confidence region is completely within the null, completely within the alternative or has non-empty intersection with both. Boeken et al. (2026) consider ternary tests to obtain Type-I and Type-II error control with finite-precision measurements. In sequential hypothesis testing, since a sample size is not determined apriori, at every stage, a sequential test decides between accepting the null, rejecting the null, and collecting another sample (Wald, 1945). See also, Berger (1985) for a Bayesian analogue of sequential hypothesis testing and Dass and Berger (2003) for examples of ternary tests in Bayesian testing of composite hypotheses. Garivier and Kaufmann (2021) consider testing M overlapping hypotheses in the sequential testing framework using M-ary tests by running M parallel binary tests with a restrictive notion of error. While some of the aforementioned works could be viewed as considering testing of overlapping hypotheses whereas others could be viewed as considering ternary tests for disjoint hypotheses, they only consider specialized cases and do not consider the question of existence of ternary tests that satisfy desiderata related to error-guarantees.

In causal inference, ternary tests were considered in passing by Robins et al. (2003); Zhang and Spirtes (2002). In addition, ternary tests have been proposed in the specific context of detecting colliders in causal discovery algorithms (Ramsey et al., 2006; Zhang and Spirtes, 2008) but our work provides a general analysis of ternary tests proving their existence under topological conditions and a way to construct ternary tests by combining binary tests. Inference for partially identified models has received a lot of attention in the econometrics literature, especially for estimation tasks; see Manski (1990); Imbens and Manski (2004); Romano and Shaikh (2010) and references within the survey paper Canay and Shaikh (2017). However, the simpler problem of hypothesis testing of partially identified queries has received little attention and further ternary tests have not been proposed or analyzed to the best of our knowledge.

2 PARTIAL IDENTIFICATION: NEED FOR TERNARY TESTS

We formalize the need for ternary tests, as mentioned in Section 1, under the semantic framework of structural causal models (SCMs) and provide definitions related to causal models in Appendix A.1. Consider a family of causal models $\mathbb{M}$ and a binary causal query, $Q : \mathbb{M} \to \{0, 1\}$. Define the causal null hypothesis

$$H_Q^0 \triangleq \{M \in \mathbb{M} : Q(M) = 0\}, \tag{1}$$

where the causal alternative hypothesis is defined by $H_Q^1 \triangleq \mathbb{M} \setminus H_Q^0$. Given that we often only have access to observational data, i.e., data sampled from $P_M(X_V)$, the set of observational distributions induced by the causal null hypothesis is given by

$$\mathcal{P}_{H_Q^0} \triangleq \left\{P_M : M \in H_Q^0\right\},$$

and the set of observational distributions induced by the causal alternative hypothesis is given by $\mathcal{P}_{H_Q^1} \triangleq \left\{P_M : M \in H_Q^1\right\}$.

If Q is partially identifiable, then $\mathcal{P}_{H_Q^0} \cap \mathcal{P}_{H_Q^1} \neq \emptyset$, i.e., while the causal null H_Q^0 and the corresponding alternative H_Q^1 are disjoint, the resulting sets of distributions that they induce, overlap, in general. When a binary-outcome statistical test with null hypothesis $\mathcal{P}_{H_Q^0}$ and disjoint alternative

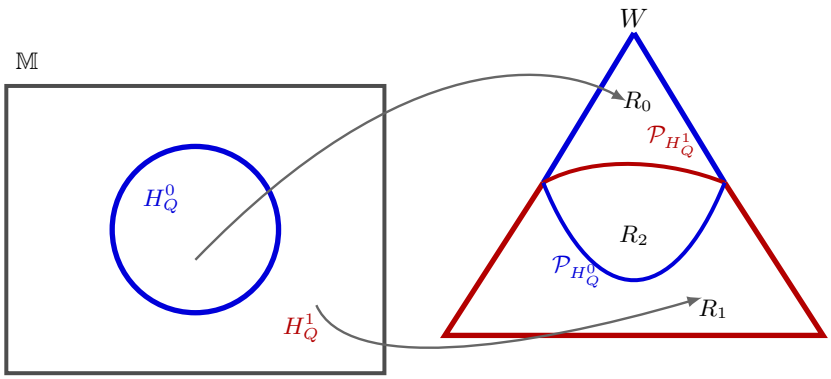


Figure 1: Illustration of a binary causal query inducing overlapping sets in the observed distributional simplex, W.

hypothesis $W \backslash \mathcal{P}_{H_Q^0}$ is used where $W \triangleq \mathcal{P}_{H_Q^0} \cup \mathcal{P}_{H_Q^1}$,[1] the results of this statistical test have the following implications for the corresponding causal hypotheses: rejecting the null implies rejecting the causal null; however, not rejecting the null does not imply that the underlying causal model is in the causal null since $\mathcal{P}_{H_Q^0} \cap \mathcal{P}_{H_Q^1} \subseteq \mathcal{P}_{H_Q^0}$ implying that given no a priori information, a distribution in $\mathcal{P}_{H_Q^0}$ could have been induced by a causal model in H_Q^1. A similar observation was also made by Zhang and Spirtes (2002) and more recently in Bhadane et al. (2025) where the relevant causal query was the absence of a direct effect. For partially identifiable causal queries, since $\mathcal{P}_{H_Q^0}$ and $\mathcal{P}_{H_Q^1}$ could overlap in general, we propose testing such queries using ternary tests. Figure 1 illustrates the need for ternary tests graphically.

A natural question that arises is how to construct a 'good' ternary test for testing a given partially identifiable binary causal query. At this stage we can decouple the context of partial identification and consider a) what are desiderata for a 'good' ternary test, and b) whether such 'good' ternary tests exist in the first place. We base our answer to the latter question on topological conditions that the statistical hypotheses must satisfy. While these topological conditions might seem restricted to the question of existence of ternary tests, we subsequently demonstrate how a practitioner could use them to guide their construction of a ternary test. Finally, we demonstrate this guide for constructing ternary tests for the partial identification examples in Section 6.

3 PRELIMINARIES

Denote $\mathcal{P}(\mathcal{X})$ as the set of all probability measures defined on the Borel σ-algebra $\mathcal{B}(\mathcal{X})$ where $\mathcal{X}$ is a separable metric space. Hypotheses $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$ are considered with respect to the weak topology on $\mathcal{P}(\mathcal{X})$ defined by the notion of weak convergence of probability measures (Billingsley, 1999). For simplicity, we assume that $W \triangleq H^0 \cup H^1$ is compact in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$ which is defined by pointwise convergence of probability measures on all Borel measurable sets, thus making it finer than the weak topology and implying that W is compact in the weak topology on $\mathcal{P}(\mathcal{X})$. See Appendix A.2 for topology-related definitions. We present a few known results below for existence of binary tests for i.i.d. data with certain desiderata of error-guarantees.

Definition 1 (Weak and Strong Consistency: Binary Tests)**.** *Given disjoint hypotheses $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$, a binary test $\phi = \{\phi_n\}_{n\in\mathbb{N}}$, where $\phi_n : \mathcal{X}^n \to \{0,1\}$ is measurable, is*

1. *weakly consistent if for all $i \in \{0,1\}$, for all $P \in H_i$, $\lim_{n\to\infty} P^n(\phi_n(X^n) = i) = 1$;*
2. *strongly consistent if for all $i \in \{0,1\}$, for all $P \in H_i$, $P^\infty(\phi_n(X^n) \neq i$ for finitely many $n) = 1$.*

Dembo and Peres (1994) provide sufficient conditions on H^0, H^1 for the existence of a strongly consistent test which are necessary if in addition the existence of $p > 1$-integrable densities for every measure in W is assumed. Under a weaker assumption of W being σ-compact, these conditions are proven to be necessary for the existence of a weakly consistent test (Ermakov, 2017, Theorem 4.4) and also for the existence of a strongly consistent test (Ermakov, 2017, Theorem 3.3). We present relevant results from Ermakov (2017) under the aforementioned assumptions on $\mathcal{X}, W$ for disjoint H^0, H^1. See Appendix D.4 for a proof.

Proposition 2. *The following are equivalent:*

(i) there exists a binary test that is weakly consistent.

(ii) H^0 and H^1 are disjoint and F_σ (i.e., countable union of closed sets) in the weak topology on $\mathcal{P}(\mathcal{X})$.

Weak and strong consistency are notions of point-wise consistency. A stronger error-guarantee desideratum is that the above hold uniformly for all P in relevant hypotheses. We

[1] Note that since $\mathcal{P}_{H_Q^1} \cap \mathcal{P}_{H_Q^0}$ is non-empty, $\mathcal{P}_{H_Q^1}$ cannot be a disjoint alternative hypothesis.

will only consider the uniform analogue of weak consistency.

Definition 3 (Uniform Consistency: Binary Tests)**.** *Given disjoint hypotheses $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$, a binary test $\phi = \{\phi_n\}_{n\in\mathbb{N}}$, where $\phi_n : \mathcal{X}^n \to \{0,1\}$ is measurable, is uniformly consistent if for all $i \in \{0,1\}$, $\lim_{n\to\infty} \sup_{P\in H_i} P^n(\phi_n(X^n) \neq i) = 0$. If the above holds only for $i = 0$, we say that ϕ is uniformly consistent with respect to H^0.*

Necessary and sufficient conditions for existence of uniformly consistent binary tests were given by Berger (1951) that are technical and difficult to apply. Le Cam and Schwartz (LeCam and Schwartz, 1960) provide the same in terms of the topology of setwise convergence on all n-fold products of probability measures, which are "rather inaccessible" (Le Cam, 1986). Much of subsequent work has focused on introducing assumptions on W so that these conditions are formulated in the weak topology. See Kleijn (2022) for a comprehensive overview. The following is a rephrasing of Ermakov (2017, Theorem 4.1) under aforementioned assumptions on $\mathcal{X}, W$ and disjoint H^0, H^1. See Appendix D.4 for a proof.

Proposition 4. *The following are equivalent:*

(i) there exists a binary test that is uniformly consistent.

(ii) H^0 and H^1 are disjoint and closed in the weak topology on $\mathcal{P}(\mathcal{X})$.

In a similar vein, Ermakov (2017) also consider existence of tests such that uniform consistency holds only for the null and weak consistency holds for the alternative (and the null since uniform consistency holds).

Proposition 5. *The following are equivalent:*

(i) there exists a weakly consistent binary test that is uniformly consistent with respect to H^0.

(ii) H^0 is closed and H^1 is F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$.

An even stronger error-guarantee desideratum on a test is that the worst-case error for H^0 (or H^1) is bounded by a predetermined constant α (or β) for every sample-size n, often referred to as Type-I (or Type-II) error guarantee[2]. Nonexistence of binary tests that control both the Type-I and Type-II error for arbitrary α or β follows from Le Cam's two-point lower bound (LeCam, 1973). Genin and Kelly (2017) consider tests with critical regions that are continuity sets (stronger than binary tests that are measurable functions) under the assumption of existence of densities for every measure in W (weaker than compactness) and provide the same topological conditions as that of Proposition 5 for existence of consistent (weakly and strongly) binary tests that control Type-I error for arbitrary α. They also provide the same topological conditions as that of Proposition 4 for existence of consistent (with respect to H^0 and H^1) ternary tests that control both the Type-I and Type-II error for arbitrary α and β. Boeken et al. (2026) consider ternary tests with open critical regions (corresponding to decisions 0 and 1) and provide the same topological conditions as that of Proposition 5 for existence of consistent binary tests that control Type-I error for arbitrary α and the same topological conditions as that of Proposition 4 for existence of consistent (with respect to H^0 and H^1) ternary tests that control both the Type-I and Type-II error for arbitrary α and β, all without making any assumptions on W.

[2]Although Boeken et al. (2026) prove that in terms of existence of tests the above notion of finite-sample error control is equivalent to uniform consistency.

4 TERNARY TESTING

As mentioned earlier, $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$ where $\mathcal{X}$ is a separable metric space and $\mathcal{P}(\mathcal{X})$ denotes the set of all Borel probability measures. In general, for overlapping hypotheses H^0 and H^1, $W \triangleq H^0 \cup H^1$ is partitioned in three disjoint sets, $R_0 \triangleq H^0 \setminus H^1$, $R_1 \triangleq H^1 \setminus H^0$ and $R_2 \triangleq H^0 \cap H^1$. Table 1a summarizes an outcome-truth table for ternary hypothesis testing. We denote errors by the outcome-truth pairs, i.e., for example, if a test outputs 0 on a sample drawn from the underlying distribution in R_2 then it makes a $(0,2)$-error. We will extend this notation to binary tests as well; i.e., a false positive is a $(1,0)$-error, and a false negative is a $(0,1)$-error. A 'ternary test' ϕ is a sequence of measurable functions $(\phi_1, \phi_2, \cdots)$ where $\phi_n : \mathcal{X}^n \to \{0,1,2\}$ with the subscript denoting the sample-size. We will abuse notation to refer to indexed families of ternary tests as ternary tests.

4.1 TESTABILITY FOR TERNARY TESTING

Given a pair of overlapping hypotheses H^0 and H^1, we will now define desiderata that ternary testing must satisfy. For binary tests, there is a long line of work that specifies such desiderata (Berger, 1951; LeCam and Schwartz, 1960; Dembo and Peres, 1994; Genin, 2018; Genin and Kelly, 2017; Ermakov, 2017; Boeken et al., 2026) in terms of consistency, uniform consistency and false positive and false negative error-control guarantees. We define analogous notions of consistency and error-control for the ternary case.

Definition 6 (Weak and Strong Consistency)**.** *Given hypotheses $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$, a ternary test $\phi = \{\phi_n\}_{n\in\mathbb{N}}$ where $\phi_n : \mathcal{X}^n \to \{0,1,2\}$ is weakly consistent if for all $i \in \{0,1,2\}$, for all $P \in R_i$, $\lim_{n\to\infty} P^n(\phi_n(X^n) = i) = 1$.*

A pair of hypotheses H^0, H^1 is weakly consistent if there

exists a test ϕ that is weakly consistent. For $P \in R_i$, we term $P^n(\phi_n(X^n) = i)$ as *probability of correct detection* (PCD). Therefore, a weakly consistent test implies that PCD approaches 1 in the limit of infinite sample-size for all $P \in R_i$, for all i. We extend the notions of error control to the ternary hypothesis testing setup below, except that errors are controlled asymptotically and not for every n.

Definition 7 ((i,j)-error control)**.** *A ternary test $\phi = \{\phi_{\alpha,n}\}_{n\in\mathbb{N},\alpha>0}$ is said to asymptotically control the (i,j)-error if for all $\alpha > 0$, $\lim_{n\to\infty} \sup_{P\in R_j} P^n(\phi_{\alpha,n}(X^n) = i) \leq \alpha$.*

In other words, controlling the (i,j)-error is an error-guarantee on concluding i when the data-generating distribution belongs to R_j. Note that if there exists a ternary test that controls the (i,j)-error then there exists a ternary test $\{\phi'_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \sup_{P\in R_j} P^n(\phi'_n(X^n) = i) = 0$ (See Section B for a proof).

Definition 8 (Ternary-Testable Hypotheses (TT(I)))**.** *For $I \subseteq \{R_0, R_1, R_2\}$, $\left(H^0, H^1\right) \subseteq \mathcal{P}(\mathcal{X})^2$ is TT(I) if there exists a weakly consistent ternary test $\phi = \{\phi_{\bar{\alpha},n}\}_{n\in\mathbb{N},\bar{\alpha}\succ 0}$ such that ϕ controls all (i,j)-errors for $i \neq j$, $R_j \in I$ where $\bar{\alpha} = \{\alpha_{ij}\}_{i\neq j, R_j\in I}$ and $\bar{\alpha} \succ 0$ implies each component is positive.*

In terms of Table 1a, a TT(I) pair of hypotheses is equivalent to the existence of a test that controls the error-rate of each of the off-diagonal cells in columns corresponding to sets in I.

To derive equivalent topological conditions on H^0, H^1 that are TT(I), we reduce the existence of ternary tests to the existence of appropriate binary tests, which implies topological conditions on hypotheses from the results listed in Section 3.The proof can be found in Appendix D.1.

Lemma 9. *If $\left(H^0, H^1\right)$ is TT(I), then all $R \in I$ are closed and all $R \notin I$ are F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$.*

In practice, the utility of these necessary topological conditions is to check whether it is possible to construct 'good' ternary tests for the notion of 'good' that the practitioner cares about. Concretely, if these topological conditions do not hold, then one can abandon the search for a good ternary test. However, if they hold, a natural question is whether these topological conditions can guide in constructing a ternary test with desired error-guarantees. We show how this can be done in the following section.

5 TWO-STAGE TERNARY TESTS: COMBINING BINARY TESTS

A natural way to obtain ternary tests is by combining two or more binary tests. Indeed, given that many binary tests have been developed and analyzed for a wide range of hypotheses, we would like to leverage existing binary tests to construct ternary tests. We term any such combination of two binary tests a "two-stage ternary test". In this section, we focus on the following specific form of two-stage ternary tests. In the first stage, we test for one of the R_i's and its complement in W[3] using a binary test. In this stage, depending on whether R_i or $R_i^{\complement}$ is the null hypothesis, we term the resulting ternary test 'Split-alternative' (SA) and 'Split-null' (SN), respectively. If in the first stage the binary test concludes to either not reject R_i or reject $R_i^{\complement}$, the ternary test concludes in favor of R_i. Otherwise, we 'split' $R_i^{\complement}$ and perform another binary test where the null and alternative hypotheses are either of the constituent sets of $R_i^{\complement}$, i.e., R_j, R_k such that $R_i^{\complement} = R_j \cup R_k$. See Figure 7 and Figure 6 in Appendix Cfor an illustration. We now relate the properties of the resulting ternary tests in terms of the properties of its constituent binary tests.

Consider two binary tests ϕ_1, ϕ_2 where the former tests the null hypothesis R_{1N} against its complement which we denote by R_{1A}, and the latter tests the null hypothesis R_{2N} against the alternative R_{2A}. Table 1c and Table 1b summarize how error-rates for errors in the outcome-truth table of Table 1a relate to the false positive ($(1,0)$-error) and false negative ($(0,1)$-error) error rates of the constituent binary tests, ϕ_1 and ϕ_2 which we denote by α_1, β_1, and α_2, β_2, respectively, i.e.,

$$\lim_{n\to\infty} \sup_{P\in R_{1N}} P^n(\phi_{1,n}(X^n) = 1) \leq \alpha_1, \tag{2}$$

$$\lim_{n\to\infty} \sup_{P\in R_{1A}} P^n(\phi_{1,n}(X^n) = 0) \leq \beta_1, \tag{3}$$

$$\lim_{n\to\infty} \sup_{P\in R_{2N}} P^n(\phi_{2,n}(X^n) = 1) \leq \alpha_2, \tag{4}$$

$$\lim_{n\to\infty} \sup_{P\in R_{2A}} P^n(\phi_{2,n}(X^n) = 0) \leq \beta_2. \tag{5}$$

The derivation of one concrete instance can be found in Appendix C.We also show that sample reuse does not affect the error-rate analysis.

5.1 COMPLETENESS OF TWO-STAGE TERNARY TESTING

So far, we related error properties of two-stage ternary tests with error properties of constituent binary tests. In this section, we show that the necessary topological conditions for a pair of hypotheses to be TT(I), i.e., necessary topological conditions for the existence of a 'good' ternary test, imply the existence of a good two-stage ternary test. Therefore, the topological conditions are both necessary and sufficient for the existence of a good ternary test.

Lemma 10. *Let $I \subseteq \{R_0, R_1, R_2\}$. If all $R \in I$ are closed and all $R \notin I$ are F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$, then*

[3]All complements are taken with respect to W.

Table 1: Outcome-truth tables for ternary tests.

Outcome \ Truth	$R_0\big(H^0 \setminus H^1\big)$	$R_1\big(H^1 \setminus H^0\big)$	$R_2\big(H^0 \cap H^1\big)$
0	True 0	$(0,1)$-error	$(0,2)$-error
1	$(1,0)$-error	True 1	$(1,2)$-error
2	$(2,0)$-error	$(2,1)$-error	True 2

(a) Outcome-truth table for a general ternary test. All off-diagonal cells are considered errors.

Outcome \ Truth	R_{1A}	R_{2N}	R_{2A}
$1A$	-	α_1	α_1
$2N$	β_1	-	β_2
$2A$	β_1	α_2	-

(b) Error table for SN test.

Outcome \ Truth	R_{1N}	R_{2N}	R_{2A}
$1N$	-	β_1	β_1
$2N$	α_1	-	β_2
$2A$	α_1	α_2	-

(c) Error table for SA test.

there exists a weakly consistent two-stage ternary test that controls all (i,j)-errors such that $i \neq j$ and $R_j \in I$.

See Appendix D.2 for a proof.From Lemma 9 and Lemma 10, we conclude that two-stage ternary testing is complete for ternary testing where errors are controlled column-wise.

Theorem 11. *Let $I \subseteq \{R_0, R_1, R_2\}$, then the following are equivalent:*

1. *The sets in I are closed and the remaining are F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$;*
2. *There exists a weakly consistent ternary test that controls all (i,j)-errors such that $i \neq j, R_j \in I$;*
3. *There exists a weakly consistent two-stage ternary test that controls all (i,j)-errors such that $i \neq j, R_j \in I$.*

Proof. $2 \implies 1 \implies 3$ from Lemma 9 and 10, respectively. $3 \implies 2$ by definition. □

5.2 GUIDELINES FOR CONSTRUCTING TWO-STAGE TERNARY TESTS

There are 12 possible two-stage ternary test options available to the practitioner, namely 6 options for the null hypothesis of the first stage and for each such option, 2 options for the second stage. Therefore, constructing a two-stage ternary test that has desired error control properties requires choosing one of the 12 options and then constructing constituent binary tests. While we do not address the latter[4], in this section, we use the topological conditions of Theorem 11 as a guide to the former part.

[4]We refer the reader to the vast literature on (binary) hypothesis testing.

When the topological conditions in Theorem 11 are satisfied for a particular I, we consider different cases depending on how many sets in $\{R_0, R_1, R_2\}$ are closed in the weak topology on $\mathcal{P}(\mathcal{X})$. Figure 8 in Appendix E illustrates decision trees that provide a guide to which two-stage ternary tests provide the desired error control under different topological properties of R_0, R_1, R_2. Note that while all the tests in the leaf nodes provide the desired error control, those suggested in blue control additional errors. While the tests highlighted in blue are suggested because they control the most number of errors, it must be noted that this excludes tests where $R_{2N} \cup R_{2A}$ is not compact because Proposition 5 depends on such a compactness assumption. As we shall see in Section 6.1, Proposition 25 can be used to guarantee existence of binary tests that are uniformly consistent w.r.t. the null even when $R_{2N} \cup R_{2A}$ is not compact. Irrespective of topological properties, for any of the aforementioned 12 options, if there exist weakly consistent binary tests for the pairs of hypotheses in the option, with error rates $\alpha_1, \beta_1, \alpha_2, \beta_2$, then the error rates of the resulting SA or SN ternary test are given in Table 1c and Table 1b, respectively.

6 APPLICATIONS TO TESTING CAUSAL QUERIES

6.1 TESTING THE INSTRUMENTAL VARIABLE (IV) INEQUALITIES

We will now apply the ternary testing framework to provide a test for the IV inequalities (Pearl, 1995). Define the IV model class, denoted by $\mathbb{M}_{\text{IV}}$, to be the set of causal models whose causal graph is a subgraph of Figure 2. Define $\mathbb{M}_{\text{IV,edge}}$ to be the set of causal models whose causal graph is a subgraph of Figure 3 and note that $\mathbb{M}_{\text{IV}} \subset \mathbb{M}_{\text{IV,edge}}$. A Markov kernel $K(X, Y \mid Z)$ defined on discrete random

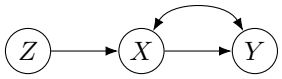


Figure 2: Causal graph (maximal) assumed for $\mathbb{M}_{\text{IV}}$

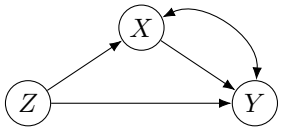


Figure 3: Causal graph (maximal) assumed for $\mathbb{M}_{\text{IV,edge}}$

variables X, Y, Z is said to satisfy the IV inequalities if

$$\max_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} \max_{z\in\mathcal{Z}} K(X=x, Y=y \mid Z=z) \leq 1. \quad (6)$$

Pearl (1995) proved that if X, Y, Z are discrete random variables, then a necessary condition for $M \in \mathbb{M}_{\text{IV}}$ is that $P_M(X, Y \mid \text{do}(Z))$ satisfies (6)[5]. We view testing the IV inequalities as testing the binary causal query that evaluates to 0 if $P_M(X, Y \mid \text{do}(Z))$ satisfies (6). We only consider the case where the instrument is binary since it is known that the IV inequalities are, in general, not sufficient for falsifying the IV model class for non-binary instruments (Bonet, 2001; Song et al., 2025). Further, Bhadane et al. (2025) show that for the binary instrument, binary outcome, categorical treatment case, the IV inequalities are sufficient.

Testing the IV inequalities is the same as testing the causal null hypothesis,

$$H^0_{\text{IV}} \triangleq \{M \in \mathbb{M}_{\text{IV,edge}} : P_M(X, Y \mid \text{do}(Z)) \text{ satisfies (6)}\}.$$

H^0_{IV} induces the set of observational distributions, $\mathcal{P}_{H^0_{\text{IV}}} \triangleq \left\{P_M : M \in H^0_{\text{IV}}\right\}$ and the causal alternative hypothesis, $H^1_{\text{IV}} \triangleq \mathbb{M}_{\text{IV,edge}} \setminus H^0_{\text{IV}}$ induces the set of observational distributions, $\mathcal{P}_{H^1_{\text{IV}}} \triangleq \left\{P_M : M \in H^1_{\text{IV}}\right\}$. As before, denote $R_0 = \mathcal{P}_{H^0_{\text{IV}}} \setminus \mathcal{P}_{H^1_{\text{IV}}}, R_1 = \mathcal{P}_{H^1_{\text{IV}}} \setminus \mathcal{P}_{H^0_{\text{IV}}}, R_2 = \mathcal{P}_{H^0_{\text{IV}}} \cap \mathcal{P}_{H^1_{\text{IV}}}$. Further, denote $W \triangleq \mathcal{P}_{H^0_{\text{IV}}} \cup \mathcal{P}_{H^1_{\text{IV}}}$ and note that for $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ being discrete and finite sets, $W = \Delta$ where Δ is the probability simplex over $|\mathcal{X}| \times |\mathcal{Y}| \times 2$.

We first characterize R_2 by showing that the distributions in R_2 are the same as those that violate positivity of Z.

Lemma 12. *Let $|\mathcal{Z}| = 2$. Then*

$$R_2 = \{P_M : P_M(z) = 0 \textit{ for some } z \in \mathcal{Z}, M \in \mathbb{M}_{\textit{IV,edge}}\}.$$

[5]Note that Pearl (1995) and most of subsequent literature frames the IV inequalities as conditions that the observational distribution $P_M(X, Y \mid Z)$ satisfies. However, this only holds under the tacit assumption of positivity which is taken for granted since the instrument, Z, is often randomized. When positivity of Z is not guaranteed, IV inequalities are correctly expressed as conditions only on $P_M(X, Y \mid \text{do}(Z))$ and thus, identifiability is not guaranteed.

Algorithm 1: Two-stage Ternary Test for (6)

Data: $(X_i, Y_i, Z_i)_{i=1}^n$
Result: $\phi_n((X_i, Y_i, Z_i)_{i=1}^n) \in \{0, 1, 2\}$
Define $f_{\text{pos}} : \mathcal{X}^n \times \mathcal{Y}^n \times \mathcal{Z}^n \to \{0, 1\}$ such that $f_{\text{pos}}((X_i, Y_i, Z_i)_{i=1}^n) = 0$ if there exists $z \in \mathcal{Z}$ such that $Z_i \neq z$ for all $1 \leq i \leq n$ and 1 otherwise;
Test R_2 vs $R_1 \cup R_0$ using ϕ_1 where $\phi_1(X_i, Y_i, Z_i)_{i=1}^n = f_{\text{pos}}((X_i, Y_i, Z_i)_{i=1}^n)$;
if ϕ_1 *outputs* 1 **then**
    Test R_0 vs R_1 using IV test implemented by Wang et al. (2017). ϕ_n outputs 1 if IV test rejects null and 0 otherwise;
else
    ϕ_n outputs 2;
end if

Algorithm 2: Two-stage Ternary Test for Treatment Efficacy Comparison

Data: $(X_i, Y_i)_{i=1}^n, c$
Result: $\phi_n((X_i, Y_i)_{i=1}^n, c) \in \{0, 1, 2\}$
Test $R_0 \cup R_2$ vs R_1 using a one-sided binomial test ϕ_1 for $P(X = 1, Y = 0)$ with threshold $1 - c$ where ϕ_1 outputs 1 if the null is rejected and 0 otherwise;
if ϕ_1 *outputs* 0 **then**
    Test R_0 vs R_2 using a one-sided binomial test ϕ_2 for $P(X = 1, Y = 1)$ with threshold c;
    ϕ_n outputs 2 if ϕ_2 rejects null and 0 otherwise;
else
    ϕ_n outputs 1;
end if

The proof can be found in Appendix D.3. We outline topological properties of R_0, R_1, R_2 and W that will eventually lead us to a two-stage ternary test for the IV inequalities.

W: Since W can be regarded as a subset of $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$, the topology of setwise convergence is equivalent to the product topology on $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|\times|\mathcal{Z}|}$. Since Δ is closed and bounded it is compact in the product topology and therefore, compact in the topology of setwise convergence.

$\mathbf{R}_2$: R_2 is closed in the weak topology on Δ since $\exists z \in \mathcal{Z}$ s.t. $P_M(Z = z) = 0$.

$\mathbf{R}_1$: For $M \in \mathbb{M}_{\text{IV,edge}}$, $P_M \in R_1$ if and only if $P_M(Z = z) > 0$ for all $z \in \mathcal{Z}$ and (6) does not hold for $P_M(X, Y \mid \text{do}(Z)) = P_M(X, Y \mid Z)$. Since the mapping $P_M(X, Y, Z) \mapsto P_M(X, Y \mid Z)$ is continuous, R_1 is open in the weak topology on Δ.

From Figure 8c, the above topological conditions suggest an SN test where $R_{1,A}$ is R_1 and $R_{2,N}$ is R_2 that controls the errors corresponding to α_1, α_2 in Table 1b, i.e., errors $(1, 0), (0, 2)$ and $(1, 2)$ are controlled. A natural candidate

for testing the IV inequalities, i.e., a positivity test followed by an IV inequalities test that assumes positivity, is given by an SA test with $R_{1N} = R_2, R_{2N} = R_0$. Since $R_{2N} \cup R_{2A}$ is not compact, Proposition 5 does not apply. However, we use Ermakov (2017)'s Theorem 4.3 (Proposition 25) to guarantee the existence of a test that asymptotically controls the $(1, 0)$ error thus implying that the SA test and the aforementioned SN test control the same errors. Algorithm 1 specifies an SA test for the IV inequalities which uses a positivity test based on the empirical distribution and the IV inequalities test assuming positivity from Wang et al. (2017).

6.2 TREATMENT EFFICACY COMPARISON

We now consider the problem of comparing the efficacy of a treatment against a threshold given data from an observational study. The treatment variable, X, has two possible states: untreated (0), treated (1). The outcome variable Y is binary indicating whether the treatment was effective (1) or not (0). Given a threshold $c \in [0, 1]$, which might have been obtained, say from a large randomized controlled trial of a different treatment, and given observational data, we want to test whether the treatment efficacy is at least the threshold, i.e., whether $P(Y = 1 \mid \text{do}(X = 1)) \geq c$.

Define $\mathbb{M}_{TEC}$ to be the set of all acyclic causal models with endogenous variables X and Y where $\mathcal{X} = \mathcal{Y} = \{0, 1\}$. The causal null hypothesis is then defined as

$$H^0_{TEC} \triangleq \{M \in \mathbb{M}_{TEC} : P_M(Y = 1 \mid \text{do}(X = 1)) \leq c\},$$

with the causal alternative hypothesis $H^1_{TEC} \triangleq \mathbb{M}_{TEC} \setminus H^0_{TEC}$. The Manski bounds (Manski, 1990) can be used to bound the above query

$$\begin{aligned} &P_M(Y = 1, X = 1) + P_M(X \neq 1) \\ &\geq P_M(Y = 1 \mid \text{do}(X = 1)) \geq P_M(Y = 1, X = 1). \quad (7) \end{aligned}$$

Since the Manski bounds are tight, the causal null hypothesis induces the set of observational distributions $\mathcal{P}_{H^0_{TEC}} \triangleq \{P_M : M \in H^0_{TEC}\}$ where

$$\mathcal{P}_{H^0_{TEC}} = \{P_M : P_M(Y = 1, X = 1) \leq c.\}$$

The causal alternative hypothesis induces the set of observational distributions $\mathcal{P}_{H^1_{TEC}} \triangleq \{P_M : M \in H^1_{TEC}\}$

$$\begin{aligned} \mathcal{P}_{H^1_{TEC}} &= \{P_M : P_M(Y = 1, X = 1) + P_M(X \neq 1) > c\}, \\ &= \{P_M : P_M(Y = 0, X = 1) < 1 - c\} \end{aligned}$$

As before, denote $R_0 = \mathcal{P}_{H^0_{TEC}} \setminus \mathcal{P}_{H^1_{TEC}}, R_1 = \mathcal{P}_{H^1_{TEC}} \setminus \mathcal{P}_{H^0_{TEC}}, R_2 = \mathcal{P}_{H^0_{TEC}} \cap \mathcal{P}_{H^1_{TEC}}$. Let $W = \mathcal{P}_{H^0_{TEC}} \cup \mathcal{P}_{H^1_{TEC}}$, and Δ be the probability simplex over $|\mathcal{X}| \times |\mathcal{Y}|$. Note that $R_0 = \Delta \setminus \mathcal{P}_{H^1_{TEC}}$ and $R_1 = \Delta \setminus \mathcal{P}_{H^0_{TEC}}$ which implies that $W = \Delta$. Since W can be regarded as a subset of $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$, the topology of setwise convergence is equivalent to the product topology on $\mathbb{R}^{|\mathcal{X}| \times |\mathcal{Y}|}$ which implies that Δ is compact in the topology of setwise convergence. Further, note that R_0 is closed and R_1 is open in the weak topology on Δ. By Figure 8c, an SN test with $R_{1A} = R_1$ and $R_{2N} = R_0$ provides maximum error control. Algorithm 2 specifies an SN test for treatment efficacy comparison where both tests use one-sided binomial proportions tests and control the $(1, 0), (2, 0)$ and $(1, 2)$ errors.

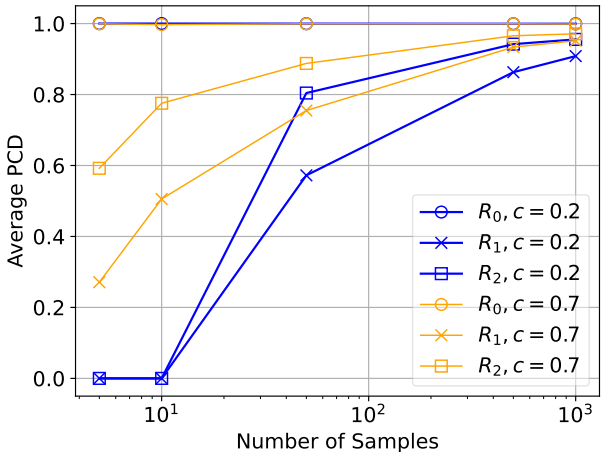


Figure 4: Average PCD of Algorithm 2 plotted as a function of number of samples. The size for the constituent binary tests was fixed at 0.025 each.

6.3 NUMERICAL EXPERIMENTS

We perform a simulation-based analysis for the treatment efficacy comparison case[6]. To make the least parametric assumptions, we consider a data-generating process where a discrete random variable U takes $|\mathcal{X}| \times |\mathcal{Y}|^{|\mathcal{X}|} = 8$ values such that each value corresponds to a response-function pair (u_x, u_y) (Balke and Pearl, 1997) where $u_x \in \{0, 1\}$ and u_y is one of four possible functions that map $\mathcal{X}$ to $\mathcal{Y}$. We sample $U \sim P_U$ and generate data $(X, Y) = (U_x, U_y(U_x))$. We sample $N_{\text{dist}} = 1000$ distributions $P_U \sim \text{Dir}(1/8)$ and further sample n i.i.d. samples of (X, Y). Figure 4 plots the average PCD of the ternary test for each of R_0, R_1 and R_2 as a function of n for two values of c, which is akin to a binary hypothesis test power analysis. Note that for R_0, for both values of c, the power is high since the SN test of Algorithm 2 controls the $(1, 0), (2, 0)$ and $(1, 2)$ errors.

To demonstrate the practical utility of the ternary testing approach, we compare the following natural confidence-interval-based ternary test (CI-based test) that a practitioner might implement: Compute level $1 - \frac{\alpha}{2}$ confidence intervals for the upper and lower Manski bounds. If the threshold c

[6]The code is at `https://github.com/SourbhBh/TernaryTesting`

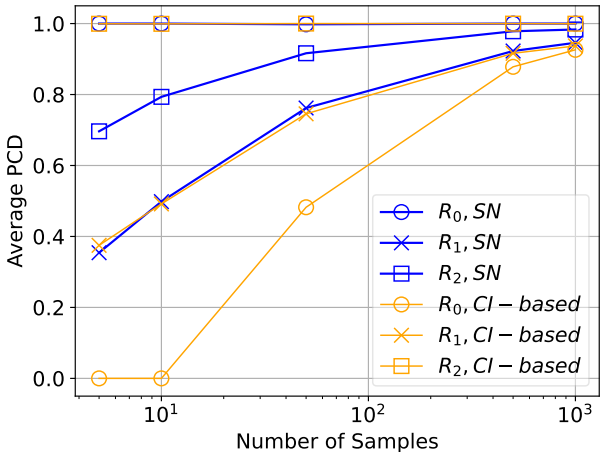


Figure 5: Comparison of the Average PCD of the CI-based ternary test versus Algorithm 2 plotted as a function of number of samples. The size for the constituent binary tests was fixed at 0.025 each and the CI-based ternary test uses 97.5% confidence intervals for the Manski bounds.

(set to 0.2) is included between the lower confidence bound of the lower Manski bound and the upper confidence bound of the upper Manski bound then the test returns 2. If the threshold is lower than the lower confidence bound of the lower Manski bound then the test returns 1 and otherwise 0. Note that this is equivalent to obtaining at least level $1-\alpha$ confidence intervals for the bound interval (Imbens and Manski, 2004). We compare our proposed SN ternary test of Algorithm 2 with the CI-based test in Figure 5. First, since both tests approach an average PCD of 1 for R_0, R_1 and R_2 both tests are weakly consistent. At smaller sample sizes, Algorithm 2 achieves a higher average PCD with respect to R_0 while the CI-based ternary test achieves a higher average PCD with respect to R_2. Depending on the decision-theoretic setup, which decide the stakes of an error, one might prefer one test over the other.

For the IV inequalities, we use the 1973 University of California, Berkeley graduate school admissions dataset `UCBAdmissions` (R Core Team, 2023) that contains counts for sex-department-admissions outcome for the 6 largest departments at University of California, Berkeley. We follow the observation of Bhadane et al. (2025) that models this decision-making process as an instrumental variable model where sex is treated as an instrument, department choice as treatment and admission decision as outcome. On the Berkeley dataset, the positivity test outputs a p-value of 0 since there is data on both sexes. The IV inequality test of Wang et al. (2017) consists of multiple directional chi-squared tests, and on the Berkeley dataset each of these satisfies the sign check and outputs 0 as the p-value of the IV inequalities test (i.e., less than the smallest positive number representable using double precision floating point format, i.e. $< 5 \times 10^{-324}$, as reported in Bhadane et al. (2025)). Thus, the IV inequalities are satisfied for the Berkeley dataset.

7 DISCUSSION

We consider ternary testing for overlapping hypotheses which includes establishing desiderata related to error-guarantees of a ternary test in terms of uniform consistency. The existence of ternary tests that satisfy these desiderata can be characterized by topological conditions on the hypotheses. We introduced two-stage ternary tests that take advantage of the large number of binary tests developed so far. Since the design space of a ternary test is arguably larger than that of binary tests, the characterizing topological conditions guide our construction of two-stage ternary tests. We demonstrate this guide on two applications, namely testing IV inequalities and treatment efficacy comparisons. In the former case, the ternary testing approach provides a theoretical grounding for the natural candidate test that first tests for positivity followed by IV tests assuming positivity. In the latter case, the ternary testing approach provides an alternate test with different error control behavior with respect to the baseline of computing confidence intervals on bounds.

We leave a) the relaxation of the compactness assumption, and b) making the categorization of ternary-testable hypotheses finer by considering existence of ternary tests that control the error-rate for any combination of cells in the error table instead of errors in columns corresponding to sets in I, for future work. As seen in the IV inequalities case and the treatment efficacy comparison study, for causal queries, the sets of interest, namely R_0, R_1 and R_2, are polynomial (linear in this case) inequalities in the observational distribution space, implying that these are semi-algebraic sets for which binary hypothesis testing has received significant attention (Drton, 2009; Sturma et al., 2024). A thorough treatment of ternary tests with constituent binary tests of semi-algebraic sets is also an interesting direction for future work.

Acknowledgements

This work was supported by Booking.com. The authors acknowledge the use of LLM-based tools for proof verifications.

REFERENCES

Balke, A. and Pearl, J. (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American statistical Association*, 92(439):1171–1176.

Berg, N. (2004). No-decision classification: an alternative

to testing for statistical significance. *the Journal of socio-Economics*, 33(5):631–650.

Berger, A. (1951). On uniformly consistent tests. *The Annals of Mathematical Statistics*, 22(2):289–293.

Berger, J. O. (1985). Statistical decision theory and Bayesian analysis. *Springer Series in Statistics*.

Bhadane, S., Mooij, J., Boeken, P., and Zoeter, O. (2025). Revisiting the Berkeley admissions data: Statistical tests for causal hypotheses. *The 41st Conference on Uncertainty in Artificial Intelligence*.

Billingsley, P. (1999). *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, NY, 2nd edition.

Boeken, P., Skapinakis, E., Genin, K., and Mooij, J. M. (2026). Topological characterisations of testable hypotheses with finite precision measurements. *arXiv preprint arXiv:2601.13946*.

Bonet, B. (2001). Instrumentality tests revisited. In *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, pages 48–55.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. (2021). Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics*, 49(5):2885–2915.

Canay, I. A. and Shaikh, A. M. (2017). Practical and theoretical advances in inference for partially identified models. *Advances in Economics and Econometrics*, 2:271–306.

Dass, S. C. and Berger, J. O. (2003). Unified conditional frequentist and Bayesian testing of composite hypotheses. *Scandinavian Journal of Statistics*, 30(1):193–210.

Dembo, A. and Peres, Y. (1994). A topological criterion for hypothesis testing. *The Annals of Statistics*, pages 106–117.

Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012.

Ermakov, M. (2017). On consistent hypothesis testing. *Journal of Mathematical Sciences*, 225(5):751–769.

Esteves, L. G., Izbicki, R., Stern, J. M., and Stern, R. B. (2016). The logical consistency of simultaneous agnostic hypothesis tests. *Entropy*, 18(7).

Garivier, A. and Kaufmann, E. (2021). Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96.

Genin, K. (2018). The topology of statistical inquiry. *PhD thesis, Carnegie Mellon University*.

Genin, K. and Kelly, K. T. (2017). The topology of statistical verifiability. *Proceedings of Theoretical Aspects of Rationality and Knowledge (TARK)*.

Hays, W. L. (1963). *Statistics for psychologists*. Holt, Rinehart and Winston, New York, N.Y., [etc.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Izbicki, R., Cabezas, L., Colugnatti, F. A., Lassance, R. F., de Souza, A. A., and Stern, R. B. (2023). REACT to NHST: Sensible conclusions to meaningful hypotheses. *arXiv preprint arXiv:2308.09112*.

Jones, L. V. and Tukey, J. W. (2000). A sensible formulation of the significance test. *Psychological methods*, 5(4):411.

Kleijn, B. J. K. (2022). *The Frequentist Theory of Bayesian Statistics*. Springer Verlag (in preparation to be published).

Le Cam, L. (1986). *Asymptotic methods in statistical decision theory*. Springer, New York.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

LeCam, L. and Schwartz, L. (1960). A necessary and sufficient condition for the existence of consistent estimates. *The Annals of Mathematical Statistics*, 31(1):140–150.

Manski, C. F. (1990). Nonparametric bounds on treatment effects. *The American Economic Review*, pages 319–323.

Munkres, J. (2000). *Topology: A First Course*. Featured Titles for Topology. Prentice Hall, Incorporated.

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in statistics-theory and methods*, 5(8):737–751.

Pearl, J. (1995). On the testability of causal models with latent and instrumental variables. In *Proceedings of the Eleventh conference on Uncertainty in Artificial Intelligence*, pages 435–443.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ramsey, J., Spirtes, P., and Zhang, J. (2006). Adjacency-faithfulness and conservative causal inference. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, pages 401–408.

Robins, J. M., Scheines, R., Spirtes, P., and Wasserman, L. (2003). Uniform consistency in causal inference. *Biometrika*, 90(3):491–515.

Romano, J. P. and Shaikh, A. M. (2010). Inference for the identified set in partially identified econometric models. *Econometrica*, 78(1):169–211.

Song, Y., Guo, F. R., Chan, K. C. G., and Richardson, T. S. (2025). The categorical instrumental variable model: Characterization, partial identification, and statistical inference. *arXiv preprint arXiv:2405.09510*.

Sturma, N., Drton, M., and Leung, D. (2024). Testing many constraints in possibly irregular models using incomplete u-statistics. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(4):987–1012.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical science*, pages 100–116.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wang, L., Robins, J. M., and Richardson, T. S. (2017). On falsification of the binary instrumental variable model. *Biometrika*, 104(1):229–236.

Zhang, J. and Spirtes, P. (2002). Strong faithfulness and uniform consistency in causal inference. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 632–639.

Zhang, J. and Spirtes, P. (2008). Detection of unfaithfulness and robust causal inference. *Minds and Machines*, 18:239–271.

Testing Partially-Identifiable Causal Queries Using Ternary Tests (Supplementary Material)

Sourbh Bhadane[1] **Joris M. Mooij**[1] **Philip Boeken**[2] **Onno Zoeter**[3]

[1]Korteweg-de Vries Institute for Mathematics, University of Amsterdam
[2]Department of Mathematics, Vrije Universiteit Amsterdam
[3]Booking.com, The Netherlands

A PRELIMINARIES

A.1 CAUSAL MODELS

We outline a few definitions related to causal models that follow the formal setup of Bongers et al. (2021).

Definition 13 (Structural Causal Model (SCM)). *A **Structural Causal Model (SCM)** is a tuple $M = (V, W, \mathcal{X}, f, P)$ where a) V, W are disjoint, finite index sets of **endogenous** and **exogenous** random variables respectively, b) $\mathcal{X} = \prod_{i \in V \cup W} \mathcal{X}_i$ is the **domain** which is a product of standard measurable spaces $\mathcal{X}_i$, c) for every $v \in V$, $f_v : \mathcal{X} \to \mathcal{X}_v$ is a measurable function, and $f = (f_v)_{v \in V}$ is called the **causal mechanism**; the equations $X_v = f_v(X)$ for every $v \in V$ are called the **structural equations**, and d) $P(\mathcal{X}_W) = \bigotimes_{w \in W} P(X_w)$ is the **exogenous distribution** which is a product of probability distributions $P(X_w)$ on $\mathcal{X}_w$.*

Definition 14 (Parent). *Let $M = (V, W, \mathcal{X}, f, P)$ be an SCM. $k \in V \cup W$ is a **parent** of $v \in V$ if and only if it is not the case that for all $x_{V \setminus k}$, $f_v(x_V, X_W)$ is constant in x_k (if $k \in V$, resp. X_k if $k \in W$) $P(\mathcal{X}_W)$-a.s..*

Definition 15 (Causal Graph). *Let $M = (V, W, \mathcal{X}, f, P)$ be an SCM. The **causal graph**, $G(M)$, is a directed mixed graph with nodes V, directed edges $u \longrightarrow v$ if and only if $u \in V$ is a parent of $v \in V$, and bidirected edges $u \leftrightarrow v$ if and only if $\exists w \in W$ that is a parent of both $u, v \in V$.*

For simplicity of exposition, we restrict attention to acyclic SCMs, i.e. SCMs whose causal graph is acyclic (contains no directed cycle $X \to \cdots \to Y \to X$).

Definition 16 (Observational Distribution). *Given an acyclic SCM, $M = (V, W, \mathcal{X}, f, P)$, the exogenous distribution, P and the causal mechanism f induce a probability distribution over the endogenous variables which is called the **observational distribution**, $P_M(X_V)$.*

Definition 17 (Hard Intervention). *Given an acyclic SCM, $M = (V, W, \mathcal{X}, f, P)$, an intervention target $T \subseteq V$, and an intervention value $x_T \in \mathcal{X}_T$, the **intervened SCM** is defined as $M_{do(X_T = x_T)} \triangleq \big(V, W, \mathcal{X}, (f_{V \setminus T}, x_T), P\big)$. Further, the observational distribution of the intervened SCM, $P_{M_{do(X_T = x_T)}}$, is called an **interventional distribution**, and denoted by $P_M(X_V \mid do(X_T = x_T))$.*

A.2 TOPOLOGY

We assume that the reader is familiar with definitions of topological spaces, metric spaces, and open and closed sets (see Munkres (2000) and Billingsley (1999)). We give brief definitions to relevant topological concepts below.

Definition 18 (Separable Metric Spaces). *A metric space $(\mathcal{X}, d)$ is separable if it contains a countable dense subset.*

The main reason to assume separability is so that the weak topology (defined below) is metrizable. Recall that $\mathcal{P}(\mathcal{X})$ is the set of all Borel probability measures defined on $\mathcal{B}(\mathcal{X})$.

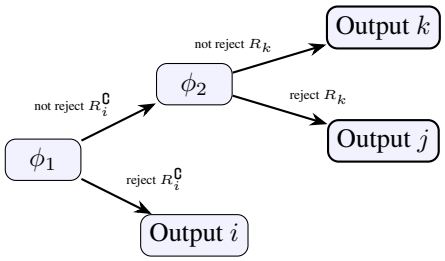


Figure 6: SN test

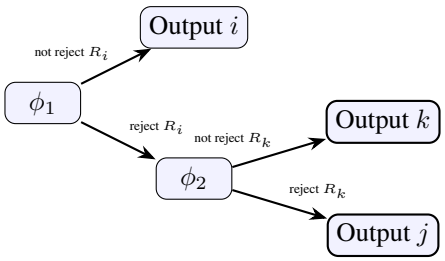


Figure 7: SA test

Definition 19 (Weak Topology)**.** *The **weak topology** on $\mathcal{P}(\mathcal{X})$ is the coarsest topology on $\mathcal{P}(\mathcal{X})$ that makes the maps $P \mapsto \int f\, dP$ continuous for every bounded continuous function $f : \mathcal{X} \to \mathbb{R}$.*

Definition 20 (Topology of Setwise Convergence)**.** *The **topology of setwise convergence** on $\mathcal{P}(\mathcal{X})$ is defined by pointwise convergence of probability measures on all Borel sets: $P_\alpha \to P$ setwise if*

$$P_\alpha(A) \longrightarrow P(A) \qquad \forall A \in \mathcal{B}(\mathcal{X}).$$

The topology of setwise convergence is defined with respect to a stronger notion of convergence of probability measures than weak convergence. This implies that every closed set of the weak topology is a closed set of the topology of setwise convergence. Consequently, there are more closed sets (and more open sets) in the latter, compared to the former, i.e., the topology of setwise convergence is **finer** than the weak topology.

Definition 21 (Compact Set)**.** *A subset $K \subseteq \mathcal{X}$ is **compact** if every open cover of K has a finite sub-cover: for every collection $\{U_\alpha\}$ of open sets with $K \subseteq \bigcup_\alpha U_\alpha$, there exist finitely many indices $\alpha_1, \ldots, \alpha_n$ such that $K \subseteq U_{\alpha_1} \cup \cdots \cup U_{\alpha_n}$.*

Clearly, compactness (defined above) in the topology of setwise convergence implies compactness in the weak topology. We use two properties of compact sets repeatedly. 1) For metric spaces (Hausdorff spaces), a set being compact implies that it is closed. 2) For Euclidean spaces, a set being closed and bounded implies that it is compact.

B ASYMPTOTIC CONTROL AND UNIFORM CONSISTENCY

Proposition 22. *Given a pair of hypotheses $H^0, H^1 \subseteq \mathcal{P}(\mathcal{X})$, if there exists a ternary test that asymptotically controls the (i,j)-error then there exists a $\phi' = \{\phi'_n\}_{n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \sup_{P\in R_j} P^n(\phi'_n(X^n) = i) = 0$.*

Proof. By Definition 7, there exists a ternary test $\phi = \{\phi_{\alpha,n}\}_{n\in\mathbb{N},\alpha>0}$ such that for all $\alpha > 0$, $\lim_{n\to\infty} \sup_{P\in R_j} P^n(\phi_{\alpha,n}(X^n) = i) \le \alpha$. Consider the sequence $\alpha_m = \frac{1}{2m}$ for $m \in \mathbb{N}$. For each $m \in \mathbb{N}$, pick $N_m \in \mathbb{N}$ such that $N_{m-1} < N_m$ (if $m > 1$) and such that for all $n > N_m$, $\sup_{P\in R_j} P^n\left(\phi_{\frac{1}{2m},n}(X^n) = i\right) < \frac{1}{m}$. Construct $\phi'_n = \phi_{\frac{1}{2m},n}$ for $N_m < n \le N_{m+1}$, and define ϕ'_n arbitrarily for $n \le N_1$. Then for any $\varepsilon > 0$, there exists $m' \in \mathbb{N}$ such that $\frac{1}{m'+1} < \varepsilon \le \frac{1}{m'}$ and consequently there exists $N_\varepsilon = N_{m'+1}$ such that for all $n > N_\varepsilon = N_{m'+1}$, n lies in some block $N_{m''} < n \le N_{m''+1}$ with $m'' \ge m' + 1$, so that

$$\sup_{P\in R_j} P^n(\phi'_n(X^n) = i) = \sup_{P\in R_j} P^n\left(\phi_{\frac{1}{2m''},n}(X^n) = i\right) < \frac{1}{m''} \le \frac{1}{m'+1} < \varepsilon.$$

This implies that $\lim_{n\to\infty} \sup_{P\in R_j} P^n(\phi'_n(X^n) = i) = 0$. □

C ERROR-RATE ANALYSIS FOR TWO-STAGE TERNARY TEST

Consider a two-stage SA ternary test with constituent binary tests ϕ_1, ϕ_2 where $R_{1N} = R_2$, $R_{2N} = R_0$. The combined ternary test is defined as

$$\phi_n(x^n) = \begin{cases} 2 & \text{if } \phi_{1,n}(x^n) = 0, \\ 0 & \text{if } \phi_{1,n}(x^n) = 1, \phi_{2,n}(x^n) = 0, \\ 1 & \text{if } \phi_{1,n}(x^n) = 1, \phi_{2,n}(x^n) = 1. \end{cases} \tag{8}$$

For simplicity, we assume that both tests use the same number of samples. We now relate the error-rates of the two-stage ternary test to the asymptotic false positive and false-negative rates of its constituent binary tests.

$(2,0)$**-error and** $(2,1)$**-error**: The error rates for these errors are given by

$$\sup_{P \in R_0} P^n(\phi_n(X^n) = 2) = \sup_{P \in R_0} P^n(\phi_{1,n}(X^n) = 0),$$
$$\sup_{P \in R_1} P^n(\phi_n(X^n) = 2) = \sup_{P \in R_1} P^n(\phi_{1,n}(X^n) = 0)$$

respectively. From (3), since $R_{1A} = R_0 \cup R_1$,

$$\max\left(\lim_{n\to\infty} \sup_{P \in R_0} P^n(\phi_n(X^n) = 2), \lim_{n\to\infty} \sup_{P \in R_1} P^n(\phi_n(X^n) = 2)\right) \leq \beta_1.$$

$(0,1)$**-error and** $(1,0)$**-error**: The error rates for these errors are given by

$$\begin{aligned} \lim_{n\to\infty} \sup_{P \in R_1} P^n(\phi_n(X^n) = 0) &= \lim_{n\to\infty} \sup_{P \in R_1} P^n(\phi_{1,n}(X^n) = 1 \cap \phi_{2,n}(X^n) = 0) \\ &\leq \lim_{n\to\infty} \sup_{P \in R_1} P^n(\phi_{2,n}(X^n) = 0) \leq \beta_2, \end{aligned}$$

$$\begin{aligned} \lim_{n\to\infty} \sup_{P \in R_0} P^n(\phi_n(X^n) = 1) &= \lim_{n\to\infty} \sup_{P \in R_0} P^n(\phi_{1,n}(X^n) = 1 \cap \phi_{2,n}(X^n) = 1) \\ &\leq \lim_{n\to\infty} \sup_{P \in R_0} P^n(\phi_{2,n}(X^n) = 1) \leq \alpha_2, \end{aligned}$$

from (5) and (4) respectively.

$(0,2)$**-error and** $(1,2)$**-error**: The error rates are given by

$$\lim_{n\to\infty} \sup_{P \in R_2} P^n(\phi_n(X^n) = 0) = \lim_{n\to\infty} \sup_{P \in R_2} P^n(\phi_{1,n}(X^n) = 1 \cap \phi_{2,n}(X^n) = 0),$$
$$\lim_{n\to\infty} \sup_{P \in R_2} P^n(\phi_n(X^n) = 1) = \lim_{n\to\infty} \sup_{P \in R_2} P^n(\phi_{1,n}(X^n) = 1 \cap \phi_{2,n}(X^n) = 1),$$

respectively. Since the right hand side of both the above equations can be bound by α_1 from (2), the result follows.

Remark: Note that the proofs above reuse samples across the two stages.

D PROOFS

D.1 PROOF OF LEMMA 9

Lemma 9. *If $\left(H^0, H^1\right)$ is TT(I), then all $R \in I$ are closed and all $R \notin I$ are F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$.*

Proof. Since $\left(H^0, H^1\right)$ are TT(I), there exists a weakly consistent ternary test $\phi = \{\phi_{\bar{\alpha},n}\}_{n\in\mathbb{N},\bar{\alpha}\succ 0}$ such that ϕ asymptotically controls all (i,j)-errors for $i \neq j$, $R_j \in I$ where $\bar{\alpha} = \{\alpha_{ij}\}_{i\neq j, R_j \in I}$.

Let $R_x \in I$. Consider the following binary test $\phi^{\text{binary}} = \{\phi^{\text{binary}}_{n,\varepsilon}\}_{n\in\mathbb{N},\varepsilon>0}$ where

$$\phi^{\text{binary}}_{n,\varepsilon}(X^n) = \begin{cases} 0 & \phi_{n,\bar{\alpha}_\varepsilon}(X^n) = x \text{ where } \bar{\alpha}_\varepsilon \triangleq \{\bar{\alpha} : \alpha_{ix} + \alpha_{kx} = \varepsilon \text{ for } i,k \neq x\} \\ 1 & \text{otherwise.} \end{cases}$$

Note that ϕ^{binary} satisfies the following: for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \sup_{P\in R_x} P^n(\phi^{\text{binary}}_{n,\varepsilon}(X^n) = 1) = \lim_{n\to\infty} \sup_{P\in R_x} P^n(\phi_{n,\bar{\alpha}_\varepsilon}(X^n)) \neq 0) \leq \varepsilon.$$

Further, for all $P \notin R_x$, $\lim_{n\to\infty} P^n(\phi^{\text{binary}}_{n,\varepsilon}(X^n) = 0) = \lim_{n\to\infty} P^n(\phi_{n,\bar{\alpha}_\varepsilon}(X^n) = 0) = 0$.

We now prove that R_x is closed in the weak topology on $\mathcal{P}(\mathcal{X})$. Assume to the contrary that R_x is not closed. This implies that there exists $P_* \in cl(R_x) \setminus R_x$ which in turn implies that $P_* \in W \setminus R_x$ since $cl(R_x) \subseteq W$ because W is compact. Consider a sequence of Borel probability measures $\{P_t\}_{t\in\mathbb{N}}$ where $P_t \in R_x$ and $P_t \to P^*$ setwise. For $\varepsilon = \frac{1}{4}$, there exists N_0 such that for all $n > N_0$, $\sup_{P\in R_x} P^n(\phi^{\text{binary}}_{n,\frac{1}{4}} = 1) < \frac{3}{8}$. By weak consistency, there exists $N > N_0$ such that $P^N_*(\phi^{\text{binary}}_{N,\frac{1}{4}} = 1) > \frac{3}{4}$. Let $A \triangleq \{x_N \in \mathcal{X}^N : \phi^{\text{binary}}_{N,\frac{1}{4}} = 1\}$ be a Borel set. We have that $P^N_t(A) \to P^N(A) > \frac{3}{4}$ since setwise convergence is preserved under finite products. Therefore, there exists t_* such that $P^N_{t_*}(A) > \frac{3}{8}$ which is a contradiction since $P_{t_*} \in R_x$. Therefore, R_x is closed in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$ implying it is closed in the weak topology on $\mathcal{P}(\mathcal{X})$.

For $R_j \notin I$, consider the following binary test $\phi^{\text{binary,j}} = \{\phi^{\text{binary,j}}_n\}_{n\in\mathbb{N}}$ where for a fixed $\bar{\alpha}$,

$$\phi^{\text{binary,j}}_n(X^n) = \begin{cases} 0 & \phi_{n,\bar{\alpha}}(X^n) = j \\ 1 & \text{otherwise.} \end{cases}$$

Since $\phi^{\text{binary,j}}$ is weakly consistent, by Proposition 2, R_j is F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$.

□

D.2 PROOF OF LEMMA 10

Lemma 10. *Let $I \subseteq \{R_0, R_1, R_2\}$. If all $R \in I$ are closed and all $R \notin I$ are F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$, then there exists a weakly consistent two-stage ternary test that controls all (i,j)-errors such that $i \neq j$ and $R_j \in I$.*

Proof. $I = \emptyset$: The statement is trivially true.

$|I| = 1$: If $R_c \in I$ is closed in the weak topology on $\mathcal{P}(\mathcal{X})$, by Proposition 5, there exists a consistent binary test that asymptotically controls the $(1,0)$-error-rate. Using this test in the first stage of an SA test where R_c corresponds to $R_{1,N}$ controls the two required errors in Table 1c. For the second stage, since the remaining sets are F_σ, by Proposition 2 there exists a consistent binary test, thus making the resulting ternary test consistent with desired error control.

$|I| = 2$: Denote the closed sets by R_{c_1} and R_{c_2}. Consider an SN test with R_{1N} as $R_{c_1} \cup R_{c_2}$. Since R_{c_1} and R_{c_2} are closed in the weak topology on $\mathcal{P}(\mathcal{X})$, their union is closed and since it is a subset of a compact set W, it is also compact. Therefore, by Proposition 4, there exists a uniformly consistent test for testing R_{c_1} as the null against R_{c_2} as the alternative (or vice versa). This implies that in Table 1b, the errors corresponding to $\alpha_1, \alpha_2, \beta_2$ are controlled. Since uniform consistency implies consistency, the existence of a consistent ternary test with desired error control is guaranteed.

$|I| = 3$: Since all the sets are closed in the weak topology on $\mathcal{P}(\mathcal{X})$, any pairwise union is closed. Since a closed subset of a compact set is compact, there exist constituent binary tests that are either an SA test or an SN test and asymptotically control both $(1,0)$-errors and $(0,1)$-errors, implying that all errors in Table 1c and Table 1b are controlled and the resulting ternary test is consistent. □

D.3 PROOF OF LEMMA 12

Lemma 12. *Let* $|\mathcal{Z}| = 2$. *Then*

$$R_2 = \{P_M : P_M(z) = 0 \textit{ for some } z \in \mathcal{Z}, M \in \mathbb{M}_{\textit{IV,edge}}\}.$$

Proof. Denote the right-hand side set in the above equation by $S_{\text{pos-viol}}$. For $M \in \mathbb{M}_{\text{IV,edge}}$, $P_M(X, Y \mid \text{do}(Z = z)) = P_M(X, Y \mid Z = z)$ if $P_M \in S^{\mathtt{C}}_{\text{pos-viol}}$. Further, in that case, $P_M(X, Y \mid \text{do}(Z))$ satisfies (6) if and only if $P_M(X, Y \mid Z)$ satisfies (6). This implies,

$$R_2 \cap S^{\mathtt{C}}_{\text{pos-viol}} = \emptyset.$$

This implies $R_2 \subseteq S_{\text{pos-viol}}$. Let $P_M \in S_{\text{pos-viol}}$ and define $z : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ as $z(x, y) \triangleq \arg\max_{z' \in \mathcal{Z}} K(x, y \mid z')$. Equation (6) can be expressed as

$$\sum_y K(x^*, y \mid z(x^*, y)) \leq 1$$

where $x^* \triangleq \arg\max_{x' \in \mathcal{X}} \sum_y K(x', y \mid z(x', y))$.

Let M be such that $P_M \in S_{\text{pos-viol}}$ and denote z' to be such that $P_M(Z = z') = 1$ and z to be such that $P_M(Z = z) = 0$.

For any M such that $P_M \in S_{\text{pos-viol}}$ and $P_M(X, Y \mid Z)$ does not satisfy (6), define M_2 such that $f_{M_2}((X, Y, Z, U_X, U_Y, U_Z, U)) = f_M((X, Y, z', U_X, U_Y, U_Z, U))$ where $z' \neq z$ and the rest of the components of the tuple defining M_2 are same as M. Therefore, $P_{M_2}(X, Y \mid \text{do}(z')) = P_{M_2}(X, Y \mid \text{do}(z))$ for $z' \neq z$. This implies that the left hand side of (6) is ≤ 1 satisfying the IV inequalities. Since $P_{M_2}(Z = z) = 0$, $P_{M_2} \in S_{\text{pos-viol}}$.

For any M such that $P_M \in S_{\text{pos-viol}}$ and $P_M(X, Y \mid Z)$ satisfies (6), choose $(\tilde{x}, \tilde{y})$ where $P_M(\tilde{x}, \tilde{y} \mid z') > 0$ and any $y^\circ \neq \tilde{y}$. Define M_2 as $f_{M_2}((X, Y, Z, U_X, U_Y, U_Z, U)) = \mathbf{1}[Z = z](\tilde{x}, y^\circ, z) + (1 - \mathbf{1}[Z = z]) f_M$. Then $P_{M_2} = P_M$, and

$$\sum_y P_{M_2}(x^*, y \mid z(x^*, y)) \geq P_M(\tilde{x}, \tilde{y} \mid z') + 1 > 1,$$

where x^* is defined with respect to M_2.

Therefore, $R_2 = S_{\text{pos-viol}}$. □

D.4 PROOFS FOR PROPOSITIONS IN SECTION 3

We prove Propositions 2, 4 and 5 by stating the version in Ermakov (2017) first in both cases. Remember that we assume $W \triangleq H^0 \cup H^1$ to be compact in the topology of setwise convergence.

Proposition 23 (Theorem 4.4 in Ermakov (2017))**.** *If* H^1 *is contained in a* σ*-compact set in the topology of setwise convergence on* $\mathcal{P}(\mathcal{X})$*, the following are equivalent:*

(i) there exists a binary test that is weakly consistent.

(ii) H^0 *and* H^1 *are contained respectively in disjoint* σ*-compact and* F_σ *(i.e., countable union of closed sets) sets respectively.*

Proposition 2. *The following are equivalent:*

(i) there exists a binary test that is weakly consistent.

(ii) H^0 *and* H^1 *are disjoint and* F_σ *(i.e., countable union of closed sets) in the weak topology on* $\mathcal{P}(\mathcal{X})$.

Proof. Since W is compact, it is closed, implying that $\bar{H^0}$ and $\bar{H^1}$ are closed subsets of W and therefore, are compact. This implies that both H^0 and H^1 are contained in σ-compact sets. If there exists a binary test that is weakly consistent, then according to Proposition 23, $H^0 \subseteq A$, $H^1 \subseteq B$ such that A is σ-compact and B is F_σ and A and B are disjoint. Then $H^1 = B \cap W$ is F_σ. Since H^0 is contained in a σ-compact set and there exists a weakly consistent binary test, according to Proposition 23, $H^0 \subseteq B_0$, $H^1 \subseteq A_0$ such that A_0 is σ-compact and B_0 is F_σ and A_0 and B_0 are disjoint. Then $H^0 = B_0 \cap W$ is F_σ. For the converse, if H^0 and H^1 are disjoint and F_σ then since H^0 is a countable union of closed sets that are subsets of a compact set, H^0 is σ-compact implying that there exists a binary test that is weakly consistent. □

Proposition 24 (Theorem 4.1 in Ermakov (2017))**.** *If H^0, H^1 are relatively compact in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$, then the following are equivalent:*

(i) there exists a binary test that is uniformly consistent.

(ii) H^0 and H^1 are such that their closures in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$ are disjoint.

Proposition 4. *The following are equivalent:*

(i) there exists a binary test that is uniformly consistent.

(ii) H^0 and H^1 are disjoint and closed in the weak topology on $\mathcal{P}(\mathcal{X})$.

Proof. Since W is compact, it is closed, implying that $\bar{H^0}$ and $\bar{H^1}$ are closed subsets of W and therefore, are compact. This implies that H^0 and H^1 are relatively compact. If there exists a binary test that is uniformly consistent, by Proposition 24, since $\bar{H^0}$ and $\bar{H^1}$ are disjoint and H^0 and H^1 are also disjoint, both H^0 and H^1 are closed in the weak topology on $\mathcal{P}(\mathcal{X})$. If H^0 and H^1 are closed in the weak topology on $\mathcal{P}(\mathcal{X})$, since W is compact implying the weak topology and the topology of setwise convergence coincide, their closures are disjoint in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$. □

Proposition 25 (Theorem 4.3 in Ermakov (2017))**.** *If H^0 is relatively compact and H^1 is contained in a set that is σ-compact in the topology of setwise convergence, then the following are equivalent:*

(i) there exists a weakly consistent binary test that is uniformly consistent with respect to H^0.

(ii) H^0 and H^1 are contained in respectively disjoint closed and F_σ sets in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$.

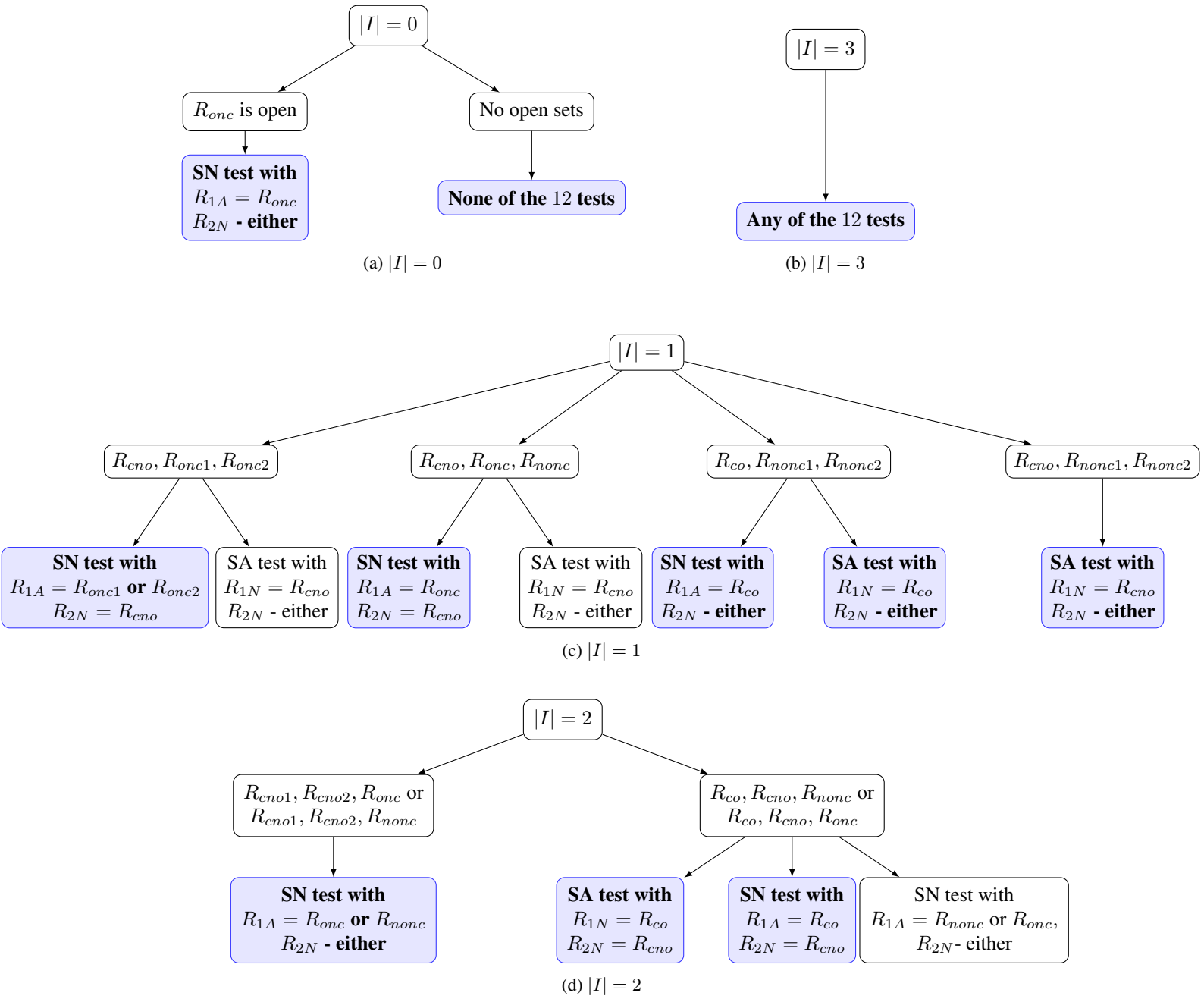
Proposition 5. *The following are equivalent:*

(i) there exists a weakly consistent binary test that is uniformly consistent with respect to H^0.

(ii) H^0 is closed and H^1 is F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$.

Proof. Since W is compact, it is closed implying that $\bar{H^0}$ and $\bar{H^1}$ are closed subsets of W and therefore, are compact. This implies that H^0 and H^1 are relatively compact. Further, $H^1 \subset W$ which is compact and also σ-compact. If (i) holds, by Proposition 25, $H^0 \subseteq A_0$ and $H^1 \subseteq B_0$ such that A_0 is closed, B_0 is F_σ, and A_0 and B_0 are disjoint. Then $H^0 = A_0 \cap W$ is closed and $H^1 = B_0 \cap W$ is F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$. Further, if H^0 and H^1 are closed and F_σ in the weak topology on $\mathcal{P}(\mathcal{X})$, then they are contained in disjoint closed and F_σ sets in the topology of setwise convergence on $\mathcal{P}(\mathcal{X})$ since W is compact. □

E DECISION TREES FOR CONSTRUCTING TWO-STAGE TERNARY TESTS



(a) $|I| = 0$

(b) $|I| = 3$

(c) $|I| = 1$

(d) $|I| = 2$

Figure 8: Decision Trees where I represents the closed sets. Subscript of R indicates topological property: cno - closed and not open, onc - open and not closed, nonc - neither closed nor open, co - closed and open. All leaf nodes provide the desired error control expected from Theorem 11. The tests suggested by blue colored leaf nodes provide the best available error control guaranteed from topological conditions although tests that control more errors are possible.