On the power of conditional independence testing under model-X

Eugene Katsevich[1] and Aaditya Ramdas[1,2]

ekatsevi@wharton.upenn.edu, aramdas@stat.cmu.edu
Department of Statistics and Data Science, University of Pennsylvania[1]
Department of Statistics and Data Science, Carnegie Mellon University[2]
Machine Learning Department, Carnegie Mellon University[2]

October 29, 2022

Abstract

For testing conditional independence (CI) of a response Y and a predictor X given covariates Z, the recently introduced model-X (MX) framework has been the subject of active methodological research, especially in the context of MX knockoffs and their successful application to genome-wide association studies. In this paper, we study the power of MX CI tests, yielding quantitative insights into the role of machine learning and providing evidence in favor of using likelihood-based statistics in practice. Focusing on the conditional randomization test (CRT), we find that its conditional mode of inference allows us to reformulate it as testing a point null hypothesis involving the conditional distribution of X. The Neyman-Pearson lemma then implies that a likelihood-based statistic yields the most powerful CRT against a point alternative. We also obtain a related optimality result for MX knockoffs. Switching to an asymptotic framework with arbitrarily growing covariate dimension, we derive an expression for the limiting power of the CRT against local semiparametric alternatives in terms of the prediction error of the machine learning algorithm on which its test statistic is based. Finally, we exhibit a resampling-free test with uniform asymptotic Type-I error control under the assumption that *only the first two moments of X given Z are known*, a significant relaxation of the MX assumption.

1 Introduction

1.1 Conditional independence testing and the MX assumption

Given a predictor $\boldsymbol{X} \in \mathbb{R}^d$, response $\boldsymbol{Y} \in \mathbb{R}^r$, and covariate vector $\boldsymbol{Z} \in \mathbb{R}^p$ drawn from a joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}$, consider testing the hypothesis of conditional

independence (CI),

$$H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z} \quad \text{versus} \quad H_1 : \boldsymbol{Y} \not\perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z}, \tag{1}$$

using n data points

$$(X, Y, Z) \equiv \{(X_i, Y_i, Z_i)\}_{i=1,\ldots,n} \overset{\text{i.i.d.}}{\sim} \mathcal{L}. \tag{2}$$

This fundamental problem—determining whether a predictor is associated with a response after controlling for a set of covariates—is ubiquitous across the natural and social sciences. To keep an example in mind throughout the paper, consider $\boldsymbol{Y} \in \mathbb{R}^1$ cholesterol level, $\boldsymbol{X} \in \{0, 1, 2\}^{10}$ the genotypes of an individual at 10 adjacent polymorphic sites, and $\boldsymbol{Z} \in \{0, 1, 2\}^{500{,}000}$ the genotypes of the individual at other polymorphic sites across the genome. Such data (X, Y, Z) would be collected in a genome-wide association study (GWAS), with the goal of testing for association between the 10 polymorphic sites of interest and cholesterol while controlling for the other polymorphic sites (1). CI testing is also connected to causal inference: with appropriate unconfoundedness assumptions, Fisher's sharp null hypothesis of no effect of a (potentially non-binary) treatment $\boldsymbol{X}$ on an outcome $\boldsymbol{Y}$ implies conditional independence. While we do not work in a causal framework, we draw inspiration from connections to causal inference throughout.

As formalized by Shah and Peters [**Shah2018**], the problem (1) is fundamentally impossible without assumptions on the distribution $\mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, in which case no asymptotically uniformly valid test of this hypothesis can have nontrivial power against *any* alternative. In special cases, the problem is more tractable, for example if $\boldsymbol{Z}$ has discrete support, or if we were willing to make (semi)parametric assumptions on the form of $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ (henceforth "model-$Y|X$"). We will not be making such assumptions in this work. Instead, we follow the lead of Candes et al. [**CetL16**], who proposed to avoid assumptions on $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$, but assume that we have access to $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$:[1]

$$\textit{model-X (MX) assumption} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}} \text{ for some known } f^*_{\boldsymbol{X}|\boldsymbol{Z}}.^2 \tag{3}$$

Candes et al. argue that while both model-$Y|X$ and MX are strong assumptions—especially when p, d are large—in certain cases much more is known about $\boldsymbol{X}|\boldsymbol{Z}$ than about $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$. In the aforementioned GWAS example, $\boldsymbol{X}|\boldsymbol{Z}$ reflects the joint distribution of genotypes at SNPs across the genome, which is well described by hidden Markov models from population genetics [**SetC17**]. On the other hand, the distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ represents the genetic basis of a complex trait, about which much less is known. In the context of (stratified) randomized experiments, the distribution $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is the propensity function [**Imai2004**] (the analog of the propensity score for non-binary treatments

[1] Candes et al. actually require that the full joint distribution $\mathcal{L}(\boldsymbol{X}, \boldsymbol{Z})$ is known, but this is because they also test for conditional associations between $\boldsymbol{Z}$ and $\boldsymbol{Y}$. We focus only on the relationship between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$ and therefore require a weaker assumption.

[2] We implicitly assume that $\mathcal{L}$ has a density with respect to some dominating measure on $\mathbb{R}^{1+1+p}$, and that all conditional densities are well-defined almost surely. Here and throughout the paper, we identify probability distributions with their densities with respect to the appropriate dominating measure.

[**Rosenbaum1983**]) and is experimentally controlled. In general causal inference contexts, the MX assumption can be viewed as the assumption that the propensity function is known.

1.2 MX methodology and open questions

Testing CI hypotheses in the MX framework has been the subject of active methodological research. The most popular methodology is MX knockoffs [**CetL16**]. This method is based on the idea of constructing synthetic negative controls (knockoffs) for each predictor variable in a rigorous way that is based on the MX assumption; see Section 5.1 for a brief overview. Rapid progress has been made on the construction of knockoffs in various cases [**SetC17**, **Romano2019a**, **Bates2019**, **Huang2019**] and on the application of this methodology to GWAS [**SetC17**, **SetS19**]. The conditional randomization test (CRT) [**CetL16**], initially less popular than knockoffs due to its computational cost, is receiving renewed attention as computationally efficient variants are proposed, such as the holdout randomization test (HRT) [**Tansey2018**], the digital twin test [**Bates2020**], and the distilled CRT (dCRT) [**Liu2020**]. The dCRT in particular is a promising methodology because it combines good power and computational speed; we focus on this variant of the CRT is Sections 3 and 4 of this paper. We introduce the general CRT methodology next, while deferring the introduction of the dCRT to Section 3.

We start with any test statistic $T(X, Y, Z)$ measuring the association between $\boldsymbol{X}$ and $\boldsymbol{Y}$, given $\boldsymbol{Z}$. Usually, this statistic involves learning some estimate $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ based on machine learning, e.g. the magnitude of the fitted coefficient for $\boldsymbol{X}$ (when $\dim(\boldsymbol{X}) = 1$) in a cross-validated lasso [**T96**] of Y on X and Z [**CetL16**]. To calculate the distribution of T under the null hypothesis (1), first define a matrix $\widetilde{X} \in \mathbb{R}^{n\times d}$, where the ith row $\widetilde{X}_i$ is a sample from $\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{Z} = Z_i)$. In other words, for each sample i, resample X_i based on its distribution conditional on the observed covariate values Z_i in that sample. We then use these resamples to build a null distribution $T(\widetilde{X}, Y, Z)$, from which the upper quantile

$$C(Y, Z) \equiv Q_{1-\alpha}[T(\widetilde{X}, Y, Z)|Y, Z] \tag{4}$$

may be extracted (the dependence on α left implicit), where the randomness is over the resampling distribution $\widetilde{X}|Y, Z$. Finally, the CRT rejects if the original test statistic exceeds this quantile:

$$\phi_T^{\mathrm{CRT}}(X, Y, Z) \equiv \begin{cases} 1, & \text{if } T(X, Y, Z) > C(Y, Z); \\ \gamma, & \text{if } T(X, Y, Z) = C(Y, Z); \\ 0, & \text{if } T(X, Y, Z) < C(Y, Z). \end{cases} \tag{5}$$

In order to accommodate discreteness, the CRT makes a randomized decision γ when $T(X, Y, Z) = C(Y, Z)$ so that the size of the test is exactly α. In practice, the threshold $C(Y, Z)$ is approximated by computing $T(\widetilde{X}^b, Y, Z)$ for a large number B of Monte Carlo resamples $\widetilde{X}^b \sim X|Z$. For the sake of clarity, this paper considers only the "infinite-B"

version of the CRT as defined by (4) and (5). In the causal inference setting, the CRT can be viewed as a variant of Fisher's exact test for randomized experiments that incorporates strata of covariates [**Zheng2008**, **Hennessy2016**], basing inference on rerandomizing the treatment to the units.

The CI testing problem under MX has benefited from several methodological innovations, but fundamental questions regarding power and optimality have received less attention. Therefore, in this paper we address the following two primary questions:

Q1. Are there "optimal" test statistics for MX methods, in any sense?

Q2. What is the precise connection between the performance of the machine learning algorithm and the power of the resulting MX method?

To the best of our knowledge, Q1 has not been considered before, while Q2 has only been indirectly addressed in the context of lasso-based knockoffs [**Weinstein2017**, **Liu2019**, **Fan2020**, **Weinstein2020**] and CRT [**Wang2020b**, **Celentano2020**]. The present paper complements these existing works by considering arbitrary machine learning methods. We summarize our findings next.

1.3 Our contributions

We find that for the MX CI problem, the CRT is more natural to analyze; it is simpler to analyze than MX knockoffs and is applicable for testing even a single conditional independence hypothesis. Thus, we focus mainly on the CRT in the present paper. We obtain the following nontrivial answers to the questions posed above.

A1: Conditional inference leads to finite-sample optimality against point alternatives. While the composite nonparametric alternative of the CI problem (1) suggests that we cannot expect to find a uniformly most powerful test, we may still ask what is the most powerful test against a point alternative. Restricting our attention to tests valid conditionally on (Y, Z) (as the CRT is) allows us to reduce the composite null to a point null. We can therefore apply the Neyman-Pearson lemma to show (Section 2) that the optimal conditionally valid test against a point alternative $\mathcal{L}$ with $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ is the CRT based on the likelihood test statistic:

$$T^{\text{opt}}(X; Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{6}$$

The same statistic yields the most powerful one-bit p-values for MX knockoffs (Section 5). Despite the simplicity of this result, it has not been derived before and appears central to the design of powerful test statistics. Since the model for $Y|X, Z$ is unknown, this result provides our first theoretical indication of the usefulness of machine learning models to learn this distribution (Q2). A2 below gives a more quantitative answer to Q2.

A2: The prediction error of the machine learning method impacts the asymptotic efficiency of the dCRT but not its consistency. It has been widely observed that the better the machine learning method approximates $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$, the higher power the MX method will have. We put this empirical knowledge on a theoretical foundation by expressing the asymptotic power of the dCRT in terms of the prediction error of the underlying machine learning method (Section 4). In particular, we consider semiparametric alternatives of the form

$$H_1 : \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}^T\beta + g(\boldsymbol{Z}), \sigma^2). \tag{7}$$

We analyze the power of a dCRT variant that employs a separately trained estimator $\widehat{g}$ in an asymptotic regime where $d = \dim(\boldsymbol{X})$ remains fixed while $p = \dim(\boldsymbol{Z})$ grows arbitrarily with the sample size n. We find that this test is consistent no matter what $\widehat{g}$ is used, while its asymptotic power against local alternatives $\beta_n = h/\sqrt{n}$ depends on the limiting mean-squared prediction error of $\widehat{g}$ (denoted $\mathcal{E}^2$) and the limiting expected variance $\mathbb{E}[\mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}]]$ (denoted s^2). For example, if $d = 1$, the

dCRT power converges to that of normal location test under alternative $N\left(\frac{hs}{\sqrt{\sigma^2 + \mathcal{E}^2}}, 1\right)$.

This represents the first explicit quantification of the impact of machine learning prediction error on the power of an MX method.

On the way to addressing Q2, we additionally establish a third result (Section 3) that may be of independent interest:

A resampling-free second-order approximation to the dCRT is equivalent to the dCRT and controls Type-I error under weaker assumptions. It was recently pointed out that if $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is Gaussian, then the resampling distribution of the dCRT test statistic can be found in closed form without actual resampling [**Liu2020**]. Here we show that the resampling-free dCRT based on the first two moments of $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is asymptotically equivalent to the dCRT based on $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ itself. Furthermore, we show the former test has asymptotic Type-I error control under the

$$\begin{aligned} &\textit{MX(2) assumption:} \text{ the first two moments of } \boldsymbol{X}|\boldsymbol{Z} \text{ are known, i.e.} \\ &\mathbb{E}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \mu(\boldsymbol{Z}) \text{ and } \mathrm{Var}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}] = \Sigma(\boldsymbol{Z}) \text{ for known } \mu(\cdot), \Sigma(\cdot). \end{aligned} \tag{8}$$

This assumption is weaker than the full MX assumption, complementing existing work [**Huang2019**, **Barber2020**] on weakening assumptions for MX methods. It also suggests that the resampling-free dCRT may be used in place of the usual dCRT while achieving similar power and controlling Type-I error asymptotically.

These advances shed new light on the nature of the MX problem and can inform methodological design. Our results handle multivariate $\boldsymbol{X}$, arbitrarily correlated designs in the model for $\boldsymbol{X}$, and any black-box machine learning method to learn $\widehat{g}$.

Notation. Recalling equations (1) and (2), population-level variables (such as $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$) are denoted in boldface, while samples of these variables (such as X_i, Y_i, Z_i) are denoted in regular font. Note that boldface does *not* distinguish between scalars, vectors, and matrices, as it is sometimes employed. The dimensions of the object in this paper will be clear from context. All vectors are treated as column vectors. We often use uppercase symbols to denote both random variables and their realizations (for either population- or sample-level quantities), but use lowercase to denote the latter when it is important to make this distinction. We use $\mathcal{L}$ to denote the joint distribution of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$, though we sometimes use this symbol to denote the joint distribution of (X, Y, Z) as well. We use the symbol "$\equiv$" for definitions. We denote by $c_{d,1-\alpha}$ the $1-\alpha$ quantile of the χ^2_d distribution, and by $\chi^2_d(\lambda)$ the non-central χ^2 distribution with d degrees of freedom and noncentrality parameter λ.

2 The most powerful CRT against point alternatives

In this section, we seek the most powerful CRT against a point alternative. To accomplish this, we make the observation—implicit in earlier works—that the CRT is valid not just unconditionally but also conditionally on Y, Z (Section 2.1). The latter conditioning step reduces the composite null to a point null. This reduction allows us to invoke the Neyman Pearson lemma to find the most powerful test (Section 2.2). Proofs are deferred to the appendix.

2.1 CRT is conditionally valid and implicitly tests a point null

Let us first formalize the definition of a level α test of the MX CI problem. The null hypothesis is defined as the set of joint distributions compatible with conditional independence and with the assumed model for $\boldsymbol{X}|\boldsymbol{Z}$:

$$\begin{aligned}\mathscr{L}_0^{\mathrm{MX}}(f^*) &\equiv \mathscr{L}_0 \cap \mathscr{L}^{\mathrm{MX}}(f^*) \\ &\equiv \{\mathcal{L} : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}\} \cap \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = f^*_{\boldsymbol{X}|\boldsymbol{Z}}\} \\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = f_{\boldsymbol{Z}} \cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}} \cdot f_{\boldsymbol{Y}|\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}, f_{\boldsymbol{Y}|\boldsymbol{Z}}\}.\end{aligned} \tag{9}$$

A test $\phi : (\mathbb{R}^d \times \mathbb{R}^r \times \mathbb{R}^p)^n \to [0,1]$ of the MX CI problem is said to be level α if

$$\sup_{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z)] \leq \alpha. \tag{10}$$

Recall that the CRT critical value $C(Y, Z)$ is defined via conditional calibration (4). As is known to those familiar with MX, this implies that any CRT $\phi = \phi_T^{\mathrm{CRT}}$ not only has level α in the sense of definition (10) but also has level α *conditionally* on Y and Z:

$$\sup_{\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)} \mathbb{E}_{\mathcal{L}}[\phi(X, Y, Z)|Y, Z] \leq \alpha \quad \text{almost surely.} \tag{11}$$

One special property of such conditionally valid tests ϕ is that they can be viewed as testing a *point null* rather than the original *composite null* (1). To see this, we view $\phi \equiv \phi(X; Y, Z)$ as a *family* of hypothesis tests, indexed by (Y, Z), for the distribution $\mathcal{L}(X|Y, Z)$. Note that under the MX assumption,

$$\mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*) \Longrightarrow \mathcal{L}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i). \tag{12}$$

In words, fixing Y, Z at their realizations y, z and viewing only X as random, $\mathcal{L}(X|Y = y, Z = z)$ equals a fixed product distribution for any null $\mathcal{L}$. This yields a conditional point null hypothesis, with respect to which $\phi_T^{\mathrm{CRT}}(x; y, z)$ is a level-α test for almost every (y, z). Note that the observations X_i in this conditional distribution are independent *but not identically distributed* due to the different conditioning events in (12).

We emphasize that the aforementioned observations have been under the hood of MX papers, and the existence of a single null distribution from which to resample $\widetilde{X}$ is central to the very definition of the CRT. Nevertheless, we find it useful to state explicitly what has thus far been largely left implicit. Indeed, viewing the CRT through the conditional lens (11) is the starting point that allows us to bring classical theoretical tools to bear on its analysis. We start doing so by considering point alternatives below.

2.2 The most powerful conditionally valid test

Viewing the CRT as a test of a point null hypothesis, we can employ the Neyman-Pearson lemma to find the most powerful CRT (in fact, the most powerful conditionally valid test) against point alternatives. The following theorem states that the likelihood ratio with respect to the (unknown) distribution $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ is the most powerful CRT test statistic against a point alternative.

Theorem 1. *Let $\bar{\mathcal{L}} \in \mathscr{L}^{MX}(f^*)$ be an alternative distribution, with $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. The likelihood of the data (X, Y, Z) with respect to $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ is*

$$T^{\mathrm{opt}}(X, Y, Z) \equiv \prod_{i=1}^{n} \bar{f}(Y_i|X_i, Z_i). \tag{13}$$

The CRT $\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}$ based on this test statistic is the most powerful conditionally valid test of $H_0 : \mathcal{L} \in \mathscr{L}_0^{\mathrm{MX}}(f^*)$ against $H_1 : \mathcal{L} = \bar{\mathcal{L}}$, i.e.

$$\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X, Y, Z)] \leq \mathbb{E}_{\bar{\mathcal{L}}}[\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}(X, Y, Z)] \tag{14}$$

for any test ϕ satisfying the conditional validity property (11).

We leave open the question of whether $\phi_{T^{\mathrm{opt}}}^{\mathrm{CRT}}$ is also the most powerful test among not just conditionally valid tests (11) but also among marginally valid tests (10). There do at least exist marginally valid tests that are not conditionally valid.

The proof of Theorem 1 (Appendix A) is based on the reduction in Section 2.1 of the composite null to a point null by conditioning, followed by the Neyman-Pearson lemma. Note that the likelihood ratio in the model $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Y},\boldsymbol{Z})$ reduces to the likelihood in the model $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$ up to constant factors; see derivation (58). This argument has similar flavor to the theory of unbiased testing (see Lehmann and Romano [**TSH**]), where uniformly most powerful unbiased tests can be found by conditioning on sufficient statistics for nuisance parameters. Our result is also analogous to but different from Lehmann's derivation of the most powerful permutation tests using conditioning followed by the Neyman-Pearson lemma, in randomization-based causal inference (see the rejoinder of Rosenbaum's 2002 discussion paper [**Rosenbaum2002**], Section 5.10 of Lehmann (1986), now Lehmann and Romano [**TSH**]).

Inspecting the most powerful test given by Theorem 1, we find that it depends on $\bar{\mathcal{L}}$ only through $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$. This immediately yields the following corollary.

Corollary 1. *Define the composite class of alternatives*

$$\begin{aligned}\mathscr{L}_1(f^*,\bar{f}) &= \{\mathcal{L}\in\mathscr{L}_0^{\mathrm{MX}}(f^*) : \bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}\}\\ &= \{\mathcal{L} : \mathcal{L}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = f_{\boldsymbol{Z}}\cdot f^*_{\boldsymbol{X}|\boldsymbol{Z}}\cdot \bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}} \text{ for some } f_{\boldsymbol{Z}}\}.\end{aligned}$$

Among the set of conditionally valid tests (11)*, the test* $\varphi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}$ *is uniformly most powerful against* $\mathscr{L}_1(f^*,\bar{f})$*.*

Theorem 1 and Corollary 1 imply that the most powerful CRT against a point alternative is based on the test statistic defined as the measuring how well the data (X,Y,Z) fit the distribution $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$. For example, if

$$\bar{f}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = N(\boldsymbol{X}^T\beta+\boldsymbol{Z}^T\gamma,\sigma^2) \text{ for coefficients } \beta\in\mathbb{R}^d \text{ and } \gamma\in\mathbb{R}^p, \tag{15}$$

then the optimal test rejects for small values of $\|Y-X\beta-Z\gamma\|^2$. In Section 5, we establish an analogous optimality statement for MX knockoffs as well. Since the optimal test depends on the alternative distribution $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z})$, CRT and MX knockoffs implementations usually employ a machine learning step to search through the composite alternative (not unlike a likelihood ratio test) for a good approximation $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. These approximate models are then summarized in various ways to define a test statistic T. There is no consensus yet on the best test statistic to use, with some authors [**CetL16**, **SetC17**, **SetS19**] using combinations of fitted coefficients $\widehat{\beta}$ and others [**Tansey2018**, **Bates2020**] using likelihood-based test statistics. The above optimality results align more closely with the latter strategy. Theorem 1 has inspired an extension of the CRT to the sequential setting using a likelihood-based test statistic, accompanied by a similar optimality result [**Grunwald2022**]. Likelihood-based test statistics also have the advantage of avoiding ad hoc combination rules for $\widehat{\beta}\in\mathbb{R}^d$ when $d>1$. It remains to be seen whether likelihood-based or coefficient-based test statistics yield greater power in practice, but a thorough empirical comparison is beyond the scope of this work. For now, it suffices to note that, despite its simplicity, this is the first such power optimality result in the CRT literature.

Intuitively, the results of this section suggest that the more successful $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ is at approximating the true alternative $f_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$, the more powerful the corresponding CRT will be. We make this relationship precise in an asymptotic setting in Section 4. We prepare for these results in the next section by exploring an easier-to-analyze asymptotic equivalent to the CRT.

3 An asymptotic equivalent to the distilled CRT

In Section 2, we saw how to construct the optimal test against point alternatives specified by $\bar{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$. In practice, of course we do not have access to this distribution, so we usually estimate it via a statistical machine learning procedure. The goal of this section and the next is to quantitatively assess the power of the CRT as a function of the prediction error of this machine learning procedure. Specifically, we consider the power of a specific instance of the CRT (the *distilled CRT (dCRT)* [**Liu2020**]) against a set of semiparametric alternatives (Section 3.1). We prepare to assess the power of this test by showing its asymptotic equivalence to the simpler-to-analyze *MX(2) F-test* (Section 3.2), which is of independent interest due to its closed form and weaker assumptions (Section 3.3). We examine the finite-sample Type-I error control of the MX(2) F-test in numerical simulations (Section B) and put this section's results into perspective (Section 3.4) before moving on to stating the desired power results in the next section (Section 4).

3.1 Semiparametric alternatives and the distilled CRT

First, we define an asymptotic framework within which we will work in Sections 3 and 4. Following a triangular array formalization, for each $n = 1, 2, \ldots$, we have a joint law $\mathcal{L}_n$ over $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{d+r+p}$, where $d = \dim(\boldsymbol{X})$ remains fixed, $r = \dim(\boldsymbol{Y}) = 1$, and $p = \dim(\boldsymbol{Z})$ can vary arbitrarily with n. For each n, we receive n i.i.d. samples $(X, Y, Z) = \{(X_i, Y_i, Z_i)\}_{i=1}^n$ from $\mathcal{L}_n$. Note that we leave implicit the dependence on n of $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ and (X, Y, Z) to lighten the notation. In this framework, it will be useful to define the mean and variance functions

$$\mu_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \text{ and } \Sigma_n(\boldsymbol{Z}) \equiv \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]. \tag{16}$$

Now, consider a set of semiparametric (partially linear) alternatives $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ such that

$$\boldsymbol{Y} = \boldsymbol{X}^T\beta_n + g_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}; \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2),\ \sigma^2 > 0 \tag{17}$$

for $\boldsymbol{\epsilon} \perp\!\!\!\perp (\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$. Here, $\beta_n \in \mathbb{R}^d$ is a coefficient vector, $g_n : \mathbb{R}^p \to \mathbb{R}$ a general function, and $\sigma^2 > 0$ the residual variance. Of special interest are local alternatives where $\beta_n = h/\sqrt{n}$ for some $h \in \mathbb{R}^d$. We emphasize that—in this section and throughout the paper—we use the partially linear model (17) exclusively as an alternative distribution against which to assess power, rather than an additional assumption required for Type-I

error control. By Theorem 1, the most powerful test against the alternative (17) is the CRT based on the likelihood statistic

$$\begin{aligned} T_n^{\text{opt}}(X, Y, Z) &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - X_i^T\beta_n - g_n(Z_i)\right)^2\right) \\ &= \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}\left(Y_i - (X_i - \mu_n(Z_i))^T\beta_n - g_n'(Z_i)\right)^2\right), \end{aligned} \tag{18}$$

where

$$\bar{g}_n(\boldsymbol{Z}) \equiv \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}] = \mu_n(\boldsymbol{Z})^T\beta_n + g_n(\boldsymbol{Z}). \tag{19}$$

Assuming local alternatives $\beta_n = h/\sqrt{n}$ and taking a logarithm, we obtain

$$\begin{aligned} \log T_n^{\text{opt}}(X, Y, Z) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - (X_i - \mu_n(Z_i))^T h/\sqrt{n} - \bar{g}_n(Z_i)\right)^2 \\ &\approx \frac{h^T}{\sigma^2}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(Y_i - \bar{g}_n(Z_i))(X_i - \mu_n(Z_i)) + C, \end{aligned} \tag{20}$$

where C is a constant that does not depend on X and therefore does not change upon resampling.

Of course, inference based on T_n^{opt} is infeasible because the function $\bar{g}_n$ is unknown in practice. Suppose we have learned an estimate $\widehat{g}_n$ of this function, possibly in-sample. Then, the derivation (20) motivates us to base inference on the sample covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ after adjusting for $\boldsymbol{Z}$:

$$\widehat{\rho}_n(X, Y, Z) \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)). \tag{21}$$

Consider first the case $d = 1$. The CRT rejecting for large values of $|\widehat{\rho}_n|$ is an instance of the dCRT [**Liu2020**]. The idea of the dCRT (Algorithm 1) is to *distill*—usually via a machine learning regression method—the information from the high-dimensional $Z \in \mathbb{R}^{p\times n}$ about X and Y into a low-dimensional summary $D \in \mathbb{R}^{q\times n}$, where $q \ll p$. This is accomplished using a *distillation function* $d : (Y, Z) \mapsto D$. Then, the CRT is applied using a test statistic of the form $T_n(X, Y, Z) \equiv T_n^d(X, Y, D) = T_n^d(X, Y, d(Y, Z))$. For example, the CRT based on the statistic $\widehat{\rho}_n$ (21) can be expressed as the dCRT with distillation function $d_i(Y, Z) = (\widehat{g}_n(Z_i), \mu_n(Z_i))$, where $\widehat{g}_n$ is learned in-sample on (Y, Z).

Algorithm 1: The distilled conditional randomization test (dCRT)

Input: $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, distribution $f^*_{\boldsymbol{X}|\boldsymbol{Z}}$, distillation function d, test statistic T_n^d, number of resamples B

Distill information in Z about X and Y into $D \equiv d(Y, Z)$;

for $b = 1, 2, \ldots, B$ **do**

  Resample $\widetilde{X}_i^{(b)} \overset{\text{ind}}{\sim} f^*_{\boldsymbol{X}|\boldsymbol{Z}=Z_i}$, $i = 1, \ldots, n$;

end

Compute $\widehat{p} \equiv \frac{1}{B+1} \sum_{b=1}^B \mathbb{1}(T_n^d(\widetilde{X}^{(b)}, Y, D) \geq T_n^d(X, Y, D))$.

Output: dCRT p-value $\widehat{p}$.

Computational cost: One p-dimensional model fit, and drawing B resamples.

The dCRT was proposed for its computational speed: The computationally expensive distillation step is a function only of (Y, Z), so it need not be refit upon resampling $\widetilde{X}$. By contrast, the originally proposed instance of the CRT [**CetL16**] involved learning $\widehat{f}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}}$ on the entire sample (X, Y, Z), and therefore the learning procedure needed to be re-applied to each resampled dataset $(\widetilde{X}^{(b)}, Y, Z)$. The derivations (18) and (20) suggest that the dCRT is not only computationally fast, but also a natural test to consider for power against semiparametric alternatives (17). We therefore focus on this class of tests.

In preparation to study the power of the dCRT, we extend it to $d > 1$ and propose an asymptotically equivalent test that is easier to analyze.

3.2 A second-order approximation to the dCRT

Let us consider first the special case

$$\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\mu_n(\boldsymbol{Z}), \Sigma_n(\boldsymbol{Z})). \tag{22}$$

In this case, the resampling distribution of $\widehat{\rho}_n$ can be computed in closed form [**Liu2020**]:

$$\mathcal{L}_n(\sqrt{n} \cdot \widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z) = N(0, \widehat{S}_n^2), \tag{23}$$

where

$$\widehat{S}_n^2 \equiv \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i). \tag{24}$$

When $d = 1$, the dCRT based on the statistic $T_n(X, Y, Z) = |\sqrt{n} \cdot \widehat{\rho}_n(X, Y, Z)|$ (and infinitely many resamples B) therefore rejects when $T_n(X, Y, Z) > \widehat{S}_n \cdot z_{1-\alpha/2}$, requiring no resampling. To extend this to $d > 1$, consider the standardized quantity

$$U_n(X, Y, Z) \equiv \widehat{S}_n^{-1} \sqrt{n} \widehat{\rho}_n = \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \in \mathbb{R}^d. \tag{25}$$

It is natural to use as a test statistic the squared norm of U_n:

$$T_n(X, Y, Z) \equiv \|U_n(X, Y, Z)\|^2. \tag{26}$$

Then, the normal resampling distribution (23) implies that

$$\mathcal{L}_n(T_n(\widetilde{X}, Y, Z)|X, Y, Z) = \chi^2_d. \tag{27}$$

It follows that the dCRT based on test statistic $T_n(X, Y, Z)$ yields the test

$$\phi_n^{N(\mu_n,\Sigma_n)}(X, Y, Z) \equiv \mathbb{1}(T_n(X, Y, Z) > c_{d,1-\alpha}), \tag{28}$$

where we recall that $c_{d,1-\alpha}$ is defined as the $1-\alpha$ quantile of χ^2_d. Note that all tests ϕ in Sections 3 and 4 will be (d)CRTs based on the test statistic T_n (26). To ease notation, we therefore omit the subscript T_n and the superscript "CRT" from the notation introduced in equation (5), replacing these with n and the distribution of $\boldsymbol{X}|\boldsymbol{Z}$ with respect to which resampling is done, respectively. For example, the superscript in the test defined in equation (28) is based on the resampling distribution $\boldsymbol{X}|\boldsymbol{Z} \sim N(\mu_n(\boldsymbol{Z}), \Sigma_n(\boldsymbol{Z}))$.

If the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is not Gaussian, then the dCRT $\phi_n^{\mathcal{L}_n}$ based on $T_n(X, Y, Z)$ will not reduce to the closed-form expression (28). However, we can think of the test $\phi_n^{N(\mu_n,\Sigma_n)}$ as a kind of second-order approximation for $\phi_n^{\mathcal{L}_n}$ as long as $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ has first and second moments given by $\mu_n(\boldsymbol{Z})$ and $\Sigma_n(\boldsymbol{Z})$, respectively. Indeed, it is easy to check that the resampling distribution $\mathcal{L}_n(\sqrt{n}\cdot\widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z)$ matches that derived in the normal case (23) up to two moments. Under a few assumptions, we can make this intuition precise by showing that $\phi_n^{\mathcal{L}_n}$ is asymptotically equivalent to $\phi_n^{N(\mu_n,\Sigma_n)}$ (Theorem 2 below). We require the distribution $\mathcal{L}_n$ to satisfy the following moment conditions [3]for fixed $c_1, c_2 > 0$:

$$\mathcal{L}_n \in \mathscr{L}_n(c_1, c_2) \equiv \{\mathcal{L}_n : \|S_n^{-1}\| \leq c_1, \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]\right] \leq c_2\}, \tag{29}$$

where

$$S_n^2 \equiv \mathbb{E}[\widehat{S}_n^2] = \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})\right]. \tag{30}$$

Furthermore, to avoid technical complications, we assume that the estimate $\widehat{g}_n$ is trained on an independent dataset (whose size can vary arbitrarily with n and is not included in the sample size n used for testing). For example, there has been recent interest in combining observational and experimental (randomized) data; typically, the former is much more abundant than the latter. We can think of $\widehat{g}_n$ being trained on the former, and then used for MX inference on the latter [**Bates2020**]. These training sets across n and resulting estimates $\widehat{g}_n$ remain fixed throughout.

Theorem 2. *Suppose that for each n, $\mathcal{L}_n$ is a law whose first and second conditional moments are given by $\mu_n(\boldsymbol{Z})$ and $\Sigma_n(\boldsymbol{Z})$ (16), which satisfies the moment conditions (29) for fixed for some $c_1, c_2 > 0$. Let $\phi_n^{\mathcal{L}_n}$ be the dCRT based on the test statistic $T_n(X, Y, Z)$ (24),*

[3]The exponents in these moment conditions can be relaxed from 4 to $2 + \delta$, in particular, requiring an appropriate triangular array weak law of large numbers with $1+\delta$ moments. This slight weakening of moment conditions requires significantly more technical effort, so is omitted for simplicity since it does not alter the main takeaway messages of our analysis.

(25), (26), *with* $\widehat{g}_n$ *trained out of sample. The threshold* $C_n(Y, Z)$ *of this test* (4) *converges in probability to the* χ^2_d *quantile:*

$$C_n(Y, Z) \xrightarrow{\mathcal{L}_n}_p c_{d,1-\alpha}. \tag{31}$$

Furthermore, if $T_n(X, Y, Z)$ *does not accumulate near* $c_{d,1-\alpha}$*, i.e.*

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = 0, \tag{32}$$

then the dCRT $\phi_n^{\mathcal{L}_n}$ *is asymptotically equivalent to its second order approximation* $\phi_n^{N(\mu_n, \Sigma_n)}$ (28)*:*

$$\lim_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathcal{L}_n}(X, Y, Z) \neq \phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z)] = 0. \tag{33}$$

Informally, this theorem (proved in Appendix C) suggests that the CRT resampling distribution of $T_n(X, Y, Z)$ converges to χ^2_d, which is the resampling distribution of this test statistic under a normal $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. Note that the resulting equivalence (33) holds for the specific instance of the CRT based on the statistic T_n defined in via equations (25) and (26), though other kinds of test statistics may lead to similar large-sample behavior. While Theorem 2 is stated for $\widehat{g}_n$ trained out of sample, we conjecture that it continues to hold when $\widehat{g}_n$ is fit in sample, as in the original dCRT construction [**Liu2020**]. At least, we observe that the conditioning in the construction of the resampling distribution $\mathcal{L}_n(\sqrt{n} \cdot \widehat{\rho}_n(\widetilde{X}, Y, Z) \mid X, Y, Z)$ ensures that its mean and variance remain equal to 0 and $\widehat{S}_n^2$ even when $\widehat{g}_n$ is fit in sample.

Theorem 2 has several consequences. First, it allows us to study the power of the dCRT $\phi_n^{\mathcal{L}_n}$ against semiparametric alternatives (17) by studying instead the simpler test $\phi_n^{N(\mu_n, \Sigma_n)}$. We pursue this direction in Section 4. Second, it implies a certain robustness property of the dCRT. Indeed, suppose we run the dCRT based on an incorrect law $\mathcal{L}'_n \neq \mathcal{L}_n$, but whose first and second moments match that of $\mathcal{L}_n$ and such that $\mathcal{L}_n$ is contiguous with respect to $\mathcal{L}'_n$. Then, applying Theorem 2 to $\mathcal{L}_n$ and $\mathcal{L}'_n$ implies that $\mathbb{P}_{\mathcal{L}_n}[\phi_n^{\mathcal{L}'_n}(X, Y, Z) \neq \phi_n^{\mathcal{L}_n}(X, Y, Z)] \to 0$. It follows that since $\phi_n^{\mathcal{L}_n}$ controls the type-I error asymptotically (in fact, also in finite samples), then so does $\phi_n^{\mathcal{L}'_n}$. We omit the formal statement of this result for the sake of brevity. Third, it suggests a distinct conditional independence test with valid Type-I error control under the weaker assumption that only the first two moments of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ are known. We expand on this third consequence next.

3.3 The MX(2) assumption and the MX(2) F-test

The asymptotic equivalence of $\phi_n^{N(\mu_n, \Sigma_n)}$ to $\phi_n^{\mathcal{L}_n}$ stated in Theorem 2 suggests that we may replace the dCRT based on the law $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ with that based on its normal approximation $N(\mu_n(\boldsymbol{Z}), \Sigma_n(\boldsymbol{Z}))$ while preserving Type-I error control. Since the test $\phi_n^{N(\mu_n, \Sigma_n)}$ requires knowledge only of the first two moments $\mu_n(\boldsymbol{X})$ and $\Sigma_n(\boldsymbol{Z})$, this means that we may control Type-I error without the full MX assumption. To formalize this, let us define the

$$\begin{gathered} \textit{MX(2) assumption:} \text{ the conditional mean } \mu_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \\ \text{and conditional variance } \Sigma_n(\boldsymbol{Z}) \equiv \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \text{ are known.} \end{gathered} \tag{34}$$

By analogy with definition (9), the MX(2) null hypothesis is defined as

$$\mathscr{L}_0^{\mathrm{MX}(2)} = \mathscr{L}_0^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \equiv \mathscr{L}_0 \cap \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)), \tag{35}$$

where

$$\mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \equiv \{\mathcal{L}_n : \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = \mu_n(\boldsymbol{Z}),\ \mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = \Sigma_n(\boldsymbol{Z})\}.$$

Under the MX(2) assumption, the CRT is undefined because there is no conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ to resample from. Nevertheless, we may define the *MX(2) F-test* by running the resampling-free dCRT as though $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ were normal, with the given first and second moments (Algorithm 2). We denote this test $\phi_n^{N(\mu_n,\Sigma_n)}$, as before.

Algorithm 2: The MX(2) *F*-test

Data: $\{(X_i, Y_i, Z_i)\}_{i=1}^n$, $\mu_n(\cdot)$ and $\Sigma_n(\cdot)$ in (16), learning method g

1 Obtain $\widehat{g}_n$ by fitting g out of sample;

2 Recall $\mu_n(Z_i) \equiv \mathbb{E}_{\mathcal{L}_n}[X_i|Z_i]$, set $\widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))^2 \Sigma_n(Z_i)$;

3 Set $U_n \equiv \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n (Y_i - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i))$ and $T_n = \|U_n\|^2$;

4 Compute $\widehat{p}^{\mathrm{MX}(2)} \equiv \mathbb{P}[\chi_d^2 > T_n]$.

Output: MX(2) F-test asymptotic p-value $\widehat{p}^{\mathrm{MX}(2)}$.

Computational cost: One p-dimensional model fit.

Note that a one-sided version of this test (the *MX(2) t-test*) can be defined for $d = 1$ by rejecting for large values of $U_n(X, Y, Z)$.

The MX(2) F-test controls the Type-I error under the MX(2) assumption, if the moment conditions (29) hold and $\widehat{g}_n$ is fit out of sample.

Theorem 3. *If $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$ for some $c_1, c_2 > 0$ and $\widehat{g}_n$ is fit out of sample, then the standardized quantity $U_n(X, Y, Z)$ converges to the standard normal:*

$$U_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{36}$$

Therefore, the MX(2) F-test controls Type-I error asympotically, uniformly over the above subset of $\mathscr{L}_0^{\mathrm{MX}(2)}$:

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1,c_2)} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n,\Sigma_n)}(X, Y, Z)] \leq \alpha. \tag{37}$$

See Appendix C for a proof of this theorem. The moment assumptions can be relaxed for pointwise error control (less desirable), but are unavoidable for uniform type-I error control as stated in the corollary. More importantly, we conjecture that the MX(2) F-test continues to have asymptotic Type-I error control even if $\widehat{g}_n$ is fit in sample. One may expect this because the validity of the MX(2) F-test derives from the correctness of (μ_n, Σ_n) rather than that of $\widehat{g}_n$. This conjecture is supported by the results of a simulation study presented in Appendix B.

3.4 Comparison to existing results

Comparison to model-X literature. The preceding results suggest that the MX(2) F-test is a useful alternative to the dCRT: the power of these methods is asymptotically the same (Theorem 2), while the MX(2) F-test is computationally faster because it does not require resampling (Table 1). On the other hand, note that we have proven Type-I error control for the MX(2) F-test only when $\widehat{g}_n$ is fit out of sample and only asymptotically, while the dCRT gives finite-sample Type-I error control with in-sample fit $\widehat{g}_n$ (albeit under the stronger model-X assumption). However, numerical simulations suggest good finite-sample Type-I error control for the MX(2) F-test even when $\widehat{g}_n$ is fit in sample. Furthermore, Theorem 3 shows that asymptotic Type-I error control of MX-style methodologies can be achieved under the weaker MX(2) assumption (34), requiring only two moments of the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ rather than the entire conditional distribution (Table 2). If strict Type-I error control in finite samples is desired, however, then we must continue to rely on the full MX assumption. Finally, note that the MX(2) assumption still requires *exact* knowledge of the first and second conditional moments; we leave as an important future direction to examine the robustness of these tests to errors in these quantities. First steps in this direction have been taken recently [**Berrett2019**, **Li2022a**].

Comparison to doubly-robust literature. The semiparametric model (17) has been extensively studied (see e.g. the classic works [**Robinson1988**, **Robins1992**]), in which context estimation of the parameter β_n is well understood. By contrast, we do not assume the validity of the semiparametric model, using it only as an alternative against which to evaluate power. A related and perhaps more relevant line of work is non-parametric doubly robust testing [**Shah2018**, **Dukes2020**] and estimation [**VanderLaan2011**, **Chernozhukov2018**]. Here, the inferential target is some functional of the data-generating distribution. The most relevant such functional is the expected conditional covariance $\rho_n \equiv \mathbb{E}_{\mathcal{L}_n}[\mathrm{Cov}_{\mathcal{L}_n}[\boldsymbol{X}, \boldsymbol{Y}|\boldsymbol{Z}]]$. Note that a valid test of the null hypothesis $H_0 : \rho_n = 0$ is also a valid test of the conditional independence hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$, since conditional independence implies that $\rho_n = 0$ (though the converse is not true in general). The quantity $\widehat{\rho}_n$ turns out to be a consistent estimator of ρ_n under the MX(2) assumption (Lemma 4). Such product-of-residuals estimators are also commonly employed in the semi- and non-parametric literatures [**Robinson1988**, **Robins1992**, **Li2011**].

To compare our results with those in non-parametric doubly robust inference, we consider the closest representative of the latter: the generalized covariance measure (GCM) test of Shah and Peters [**Shah2018**]. For $d = 1$, the GCM test statistic is defined as

$$\widehat{\rho}_n^{\mathrm{GCM}} \equiv \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))(X_i - \widehat{\mu}_n(Z_i)), \tag{38}$$

where $\widehat{\mu}_n(\boldsymbol{Z})$ and $\widehat{g}_n(\boldsymbol{Z})$ are estimates of $\mu_n(\boldsymbol{Z}) = \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $g_n(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$, respectively. This statistic is shown to converge under conditional independence to a mean-zero normal limit as long as the estimates of $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$ are both

consistent, while the product in these estimation errors tends to zero at a rate of $o(n^{-1/2})$. By contrast, the MX(2) F-test places more weight on the model for $\boldsymbol{X}|\boldsymbol{Z}$ (assuming both first and second moments of this conditional distribution are known) while placing less weight on the model for $\boldsymbol{Y}|\boldsymbol{Z}$ (not assuming even consistency for $\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]$). Therefore, while the MX(2) F-test closely resembles the GCM test, the assumptions required for validity of these two methods do not subsume each other (Table 2).

Method	Guarantee	Resampling
CRT	Finite-sample	Yes
MX(2) F-test	Asymptotic	No
GCM test	Asymptotic	No

Table 1: Type-I error guarantee and necessity of resampling for each method compared.

Method	$\mathcal{E}(\mathbb{E}[\boldsymbol{X}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathrm{Var}[\boldsymbol{X}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathcal{L}(\boldsymbol{X}\|\boldsymbol{Z}))$	$\mathcal{E}(\mathbb{E}[\boldsymbol{Y}\|\boldsymbol{Z}])$	$\mathcal{E}(\mathbb{E}[\boldsymbol{X}\|\boldsymbol{Z}]) \times \mathcal{E}(\mathbb{E}[\boldsymbol{Y}\|\boldsymbol{Z}])$
CRT	0	0	0	–	–
MX(2) F-test	0	0	–	–	–
GCM test	$o_p(1)$	–	–	$o_p(1)$	$o_p(n^{-1/2})$

Table 2: Assumptions necessary for each method compared (excluding moment assumptions). Here, $\mathcal{E}(\cdot)$ refers to the root-mean-squared estimation error of a given quantity.

Comparison to causal inference literature. Theorem 2 is a statement about the asymptotic equivalence between the resampling-based CRT and the asymptotic MX(2) F-test. The MX CRT is in the spirit of the finite-population approach to causal inference (Fisher), whereas the MX(2) F-test is in the spirit of the asymptotic super-population approach (Neyman). We find that research in these two strands of work on causal inference have proceeded largely separately from each other, and therefore connections between the two have received relatively little attention. However, there has been a recent line of work [**Ding2017**, **Wu2020a**, **Zhao2021**] focusing on the asymptotic behavior of the Fisher randomization test in the context of completely randomized experiments. A similar result to Theorem 2 is that the Fisher randomization test (analogous to the CRT) is asymptotically equivalent to the Rao score test (analogous to the MX(2) F-test) in a completely randomized experiment [**Ding2017**]. Theorem 2 can be viewed as an extension of this result to accommodate for non-binary treatments as well as high-dimensional covariates affecting both treatment and response.

Having found that the dCRT is a natural test to apply for power against semiparametric alternatives, and that this test is equivalent to the simpler MX(2) F-test, we are ready to study the relationship between the power of the dCRT and the quality of the underlying machine learning procedure.

4 dCRT power against semiparametric alternatives

In this section, we present our results on the asymptotic power of the dCRT against the semiparametric alternatives (17). We state these results first (Section 4.1), then apply these to lasso-based dCRT (Section 4.2), and finally compare our results to existing ones (Section 4.3). All proofs are deferred to Appendix C.

4.1 Power against semiparametric alternatives

In Theorem 4 below, we express the asymptotic power of the dCRT against alternatives (17) in terms of the variance-weighted mean square error of $\widehat{g}_n$:

$$\mathcal{E}_n^2 \equiv \mathbb{E}_{\mathcal{L}_n}\left[(\widehat{g}_n(\boldsymbol{Z}) - \bar{g}_n(\boldsymbol{Z}))^2 \cdot \overline{\Sigma}_n^{-1/2}\Sigma_n(\boldsymbol{Z})\overline{\Sigma}_n^{-1/2}\right], \quad \text{where } \overline{\Sigma}_n \equiv \mathbb{E}_{\mathcal{L}_n}[\Sigma_n(\boldsymbol{Z})]. \tag{39}$$

Recall from definition (19) that $\bar{g}_n(\boldsymbol{Z}) \equiv \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$. Note that if $(\boldsymbol{X}, \boldsymbol{Z})$ is jointly Gaussian, then $\Sigma_n(\boldsymbol{Z}) = \overline{\Sigma}_n$ for all $\boldsymbol{Z}$ and therefore $\mathcal{E}_n^2 = \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - \bar{g}_n(\boldsymbol{Z}))^2] \cdot I_d$. Our result requires the following moment assumptions:

$$\sup_n \|\overline{\Sigma}_n^{-1}\| < \infty, \tag{40}$$

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^8] < \infty, \tag{41}$$

and

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - \bar{g}_n(\boldsymbol{Z}))^4\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] < \infty. \tag{42}$$

Theorem 4. *Suppose $\mathcal{L}_n$ and $\widehat{g}_n$ (trained out of sample) are such that the conditional distribution $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ follows the semiparametric alternative* (17)*, the moment conditions* (40)*,* (41)*, and* (42) *are satisfied, and that the conditional variance and variance-weighted mean squared error converge:*

$$\overline{\Sigma}_n \to \overline{\Sigma} \quad \textit{and} \quad \mathcal{E}_n^2 \to \mathcal{E}^2 \quad \textit{as } n \to \infty. \tag{43}$$

Then, we have the following two statements:

(a) (Consistency) If $\beta_n = \beta \neq 0$ for each n, then the dCRT $\phi_n^{\mathcal{L}_n}$ and the MX(2) F-test $\phi_n^{N(\mu_n,\Sigma_n)}$ are consistent:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathcal{L}_n}(X, Y, Z)\right] = \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{N(\mu_n,\Sigma_n)}(X, Y, Z)\right] = 1. \tag{44}$$

(b) (Power against local alternatives) If $\beta_n = h_n/\sqrt{n}$ for a convergent sequence $h_n \to h \in \mathbb{R}^d$, then

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{\mathcal{L}_n}(X, Y, Z)\right] &= \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}\left[\phi_n^{N(\mu_n,\Sigma_n)}(X, Y, Z)\right] \\ &= \mathbb{P}[\chi_d^2(\|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2}h\|^2) > c_{d,1-\alpha}].\end{aligned} \tag{45}$$

This theorem is proved in Appendix C. Recalling that $\chi^2_d(\lambda)$ denotes the noncentral chi-square distribution with d degrees of freedom and non-centrality parameter λ and $c_{d,1-\alpha}$ denotes the $1-\alpha$ quantile of χ^2_d, the second part of Theorem 4 states that the dCRT has power equal to that of a χ^2 test of a multivariate normal random vector having mean zero under the alternative $N((\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2}h, I_d)$. This result establishes a direct link between the estimation error in $\widehat{g}_n$ and the power of the CRT against local alternatives. In particular, the mean-squared error term $\mathcal{E}^2$ contributes additively to the irreducible error term $\sigma^2 I_d$. We can gain intuition for this result by considering the regression model

$$\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}) = (\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n + (\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}) \tag{46}$$

obtained from the semiparametric model (17) by subtracting $\widehat{g}_n(\boldsymbol{Z})$ from both sides. The test statistic T_n is based on the quantity $\widehat{\rho}_n$ defined in equation (21), which can be viewed as an unnormalized version of the fitted regression coefficients of $Y - \widehat{g}_n(Z)$ on $X - \mu_n(Z)$. The term $\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z})$ in the regression model (46) contributes additively to the residual error term, so in a traditional regression analysis we would expect the power of the test to depend on the variance of this error term. In fact, standard large-sample OLS theory (see e.g. Section 2.3 of Hayashi's book [**Hayashi2000**]) states that the power against local alternatives of the F-test in the regression model (46) is exactly the same as that of the dCRT (and MX(2) F-test) stated in equation (45). Of course, the usual F-test applied to the regression (46) relies on the validity of this model while the dCRT and MX(2) F-test instead rely on knowledge of $\mathrm{Var}[\boldsymbol{X}|\boldsymbol{Z}]$. Note that [**Wang2020b**] also find the power of an MX test and a classical OLS test to have the same power (see their Appendix F).

4.2 Example: Power of lasso-based CRT

A key ingredient in the power formula (45) is the limiting variance-weighted mean squared error $\mathcal{E}^2$. This error depends on the machine learning method used to obtain $\widehat{g}_n$. We can leverage existing results about the asymptotic behavior of prediction error of machine learning methods in high dimensions. In this section, we consider the case when $\widehat{g}_n$ is trained using the lasso in the orthogonal design case, which was studied by Bayati and Montanari [**Bayati2011**]. Note that a recent extension of Bayati and Montanari's results to correlated designs [**Celentano2020**] can also be used in tandem with (45), but we focus our exposition on the orthogonal design case for the sake of simplicity.

Setting 1 (Linear regression with orthogonal design). Consider a sequence of laws $\mathcal{L}_n$ such that $\mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Z}) = N(0, I_{1+p})$ and such that $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ follows the semiparametric model (17), with $\beta_n = h_n/\sqrt{n}$ for some convergent sequence $h_n \to h \in \mathbb{R}$ and $g_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\gamma_n$ for a sequence $\gamma_n \in \mathbb{R}^p$ such that the entries of $\sqrt{n}\gamma_n$ converge weakly to a random variable Γ on $\mathbb{R}$ with $\mathbb{P}[\Gamma \neq 0] > 0$ and $\|\sqrt{n}\gamma_n\|^2/p \to \mathbb{E}[\Gamma^2] < \infty$.

Until now, we have denoted by n the sample size used for constructing tests, leaving unspecified the size of the separate sample used to train $\widehat{g}_n$. To get concrete expressions

for the power of the dCRT based on a specific machine learning method to obtain $\widehat{g}_n$, we must take the training sample size into account, which we will do via sample splitting for convenience. We therefore define the test $\varphi_n^{\mathcal{L}_n}(X, Y, Z)$, which for some training proportion $\pi \in (0, 1)$ split the data into πn training observations $(X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}})$ and $(1-\pi)n$ test observations $(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}})$. This test proceeds by first running a lasso of Y_{train} on Z_{train} with regularization parameter λ to obtain an estimate $\widehat{\gamma}_{\pi n}$. The test $\varphi_n^{\mathcal{L}_n}(X, Y, Z)$ is then obtained by running the dCRT on the test data $(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}})$ using the estimate $\widehat{g}_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\widehat{\gamma}_{\pi n}$:

$$\varphi_n^{\mathcal{L}_n}(X, Y, Z) \equiv \phi_{(1-\pi)n}^{\mathcal{L}_n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}}).$$

Note that the dependence of $\phi_{(1-\pi)n}^{\mathcal{L}_n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}})$ on the training data $(X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}})$ is left implicit.

Under Setting 1, we can directly use Bayati and Montanari's theory [**Bayati2011**] to obtain

$$\lim_{n\to\infty} \mathcal{E}_n^2 = \tau_*^2 - \sigma^2 \quad \text{a.s. in } (X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}), \tag{47}$$

where (α_*, τ_*) is the unique solution of the system below:

$$\begin{aligned} \lambda &= \alpha\tau(1 - (\pi\delta)^{-1}\mathbb{E}[\eta'(\sqrt{\pi}\Gamma + \tau W; \alpha\tau)]), \\ \tau^2 &= \sigma^2 + (\pi\delta)^{-1}\mathbb{E}[(\eta(\sqrt{\pi}\Gamma + \tau W; \alpha\tau) - \sqrt{\pi}\Gamma)^2]. \end{aligned} \tag{48}$$

Here, $W \sim N(0, 1)$ is independent of Γ and $\eta(x; \theta) = (|x| - \theta)_+\text{sign}(x)$ is the soft threshold function. This leads to the following corollary of Theorem 4, proved in Appendix C:

Corollary 2. *Under Setting 1, the asymptotic power of the dCRT converges to that of a standard normal location test with alternative mean* $\tau_*^{-1}h\sqrt{1-\pi}$*:*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\mathcal{L}_n}(X, Y, Z)] = \mathbb{P}[|N(\tau_*^{-1}h\sqrt{1-\pi}, 1)| > z_{1-\alpha/2}]. \tag{49}$$

Corollary 2 gives the power of these lasso-based methods in a very simple form, with the prediction error of the lasso entering through the effective noise level τ_*. The impact of the splitting proportion π on power can be seen in the multiplication of the signal strength h by $\sqrt{1-\pi}$. The splitting proportion implicitly impacts the effective noise level τ_* as well; smaller π lead to greater effective noise levels. Note that the expectations in Corollary 2 are over both training and test sets, while the expectations in Theorem 4 are over the test set only.

4.3 Comparison to existing results

Two other power analyses of the CRT have been recently conducted [**Wang2020b**, **Celentano2020**] in parallel to the first version of our paper [**Katsevich2020a**], focusing on the case where $g_n(\boldsymbol{Z}) = \boldsymbol{Z}^T\gamma_n$, $\widehat{g}_n$ is trained using the lasso, $n/p \to \delta$, and the generalized covariance measure test statistic $\widehat{\rho}_n$ is used. The former study considers the case of orthogonal design (Setting 1), while the latter considers arbitrary joint Gaussian distribution for $(\boldsymbol{X}, \boldsymbol{Z})$. Assuming $\mathbb{E}_{\mathcal{L}_n}[\text{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]] \to s^2$ (the quantity we called $\overline{\Sigma}$ in

Section 4.1, with different notation to clarify that for $\dim(\boldsymbol{X}) = 1$ the covariance matrix simply becomes a variance), the works [**Wang2020b**, **Celentano2020**] found that the power of the CRT with in-sample lasso fit tends to that of a normal location test with alternative mean sh/τ_*, where τ_* is the effective noise level from AMP theory (in the orthogonal design case, (α_*, τ_*) are defined by equation (48) with $\pi = 1$) and h is the limiting constant of the local alternatives in Setting 1.

This is a similar expression to what we found in Corollary 2 in the orthogonal design case. Furthermore, note that $\tau_*^2 = \sigma^2 + \mathcal{E}^2$ (i.e. the out-of-sample prediction error of the lasso). It follows that the power expression found by [**Wang2020b**, **Celentano2020**] is exactly the same as what we found in part (b) of Theorem 4, despite the fact that their $\widehat{g}_n$ is fit in-sample. [**Wang2020b**] also derive a power expression for the CRT when $\widehat{g}_n$ is fit in-sample via ordinary least squares (allowing correlated covariates, as we do in Theorem 4), which also happens to coincide with expression (45). Such in-sample results have been obtained only for these two test statistics, though we conjecture that such results hold more broadly. By contrast, training $\widehat{g}_n$ on a separate sample allows us to prove Theorem 4 for very broad (almost unrestricted) classes of machine learning methods $\widehat{g}_n$.

Finally, we note a connection between Theorem 4 and causal inference. It is widely known in causal inference (see e.g. [**Imbens2015**]) that adjustment for covariates Z in randomized experiments (a) yields consistent estimates despite misspecification of $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$ and (b) improve estimation efficiency to the extent that this adjustment captures the distribution $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z})$. This fact mirrors the conclusions of Theorem 4. The asymptotic variance of the regression-based estimator for the average treatment effect in a completely randomized experiment is a standard result, but we are unaware of a quantitative expression of the asympotic efficiency of covariate-adjusted versions of the Fisher randomization test (though some insight is provided by [**Zhao2021**]).

5 Most powerful one-bit p-values for knockoffs

MX knockoffs [**CetL16**] operate differently than the CRT; they simultaneously test the conditional associations of many variables with a response. Given m variables $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m$ and a response $\boldsymbol{Y}$, it is of interest to test the CI hypotheses

$$H_j : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X}_j \mid \boldsymbol{X}_{-j}, \quad j = 1, \ldots, m.$$

Note that j indexes variables, rather than samples. Comparing to our setup, $\boldsymbol{X}_j$ plays the role of $\boldsymbol{X}$ and $\boldsymbol{X}_{-j}$ plays the role of $\boldsymbol{Z}$. In particular, we allow $\boldsymbol{X}_j$ to be a group of variables. Like HRT, knockoffs only requires one model fit, so it too is computationally faster than the CRT. Among these three MX procedures, knockoffs is currently the most popular. We briefly review it next, and then present an optimality result in the spirit of Theorem 1. Its proof is given in Appendix D.

5.1 A brief overview of knockoffs

A set of knockoff variables $\widetilde{\boldsymbol{X}} = (\widetilde{\boldsymbol{X}}_1, \ldots, \widetilde{\boldsymbol{X}}_m)$ is constructed to satisfy conditional exchangeability:

$$\mathcal{L}(\boldsymbol{X}_j, \widetilde{\boldsymbol{X}}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}) = \mathcal{L}(\widetilde{\boldsymbol{X}}_j, \boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}_{-j}), \quad j = 1, \ldots, m \tag{50}$$

and conditional independence

$$\boldsymbol{Y} \perp\!\!\!\perp \widetilde{\boldsymbol{X}} \mid \boldsymbol{X}. \tag{51}$$

Here, $\boldsymbol{X}_j \in \mathbb{R}$ denotes the jth element of the vector $\boldsymbol{X} \in \mathbb{R}^m$ and $\boldsymbol{X}_{-j} \in \mathbb{R}^{m-1}$ denotes all elements except the jth. Also, $X_{i,\bullet} \in \mathbb{R}^n$, $X_{\bullet,j} \in \mathbb{R}^m$, and $X_{\bullet,-j} \in \mathbb{R}^{n \times (m-1)}$ denote the ith row, jth column, and all columns but the jth of the matrix $X \in \mathbb{R}^{n \times m}$. Given such a construction, a set of knockoff variables $\widetilde{X}_{i,\bullet}$ is sampled from $\mathcal{L}(\widetilde{\boldsymbol{X}} | \boldsymbol{X} = X_{i,\bullet})$ for each i. Knockoff inference is then based on a form of data-carving: variables are given an ordering $\tau(1), \ldots, \tau(m)$ determined arbitrarily from $([X, \widetilde{X}], Y)$ as long as $X_{\bullet,j}$ and $\widetilde{X}_{\bullet,j}$ are treated symmetrically. Variables are then tested in that order based on *one-bit p-values* p_j measuring the contrast between the strength of association between $X_{\bullet,j}$ and Y and that between $\widetilde{X}_{\bullet,j}$ and Y. Given any statistic $T_j([X, \widetilde{X}], Y)$ measuring the strength of association between X_j and Y, define the one-bit p-value

$$p_j([X, \widetilde{X}], Y) \equiv \begin{cases} \frac{1}{2}, & \text{if } T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y); \\ 1, & \text{if } T_j([X, \widetilde{X}], Y) \leq T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y). \end{cases} \tag{52}$$

Here, $[X, \widetilde{X}]_{\text{swap}(j)}$ is defined as the result of swapping $X_{\bullet,j}$ with $\widetilde{X}_{\bullet,j}$ in $[X, \widetilde{X}]$ while keeping all other columns in place. A set of variables with guaranteed false discovery rate control is chosen via the ordered testing procedure *Selective SeqStep* [**BC15**], applied to the p-values p_j in the order τ.

5.2 Most powerful one-bit p-value

It is harder to analyze the power of knockoffs than that of the CRT for several reasons. Knockoffs is fundamentally a *multiple* testing procedure, coupling the analysis of H_j across variables j. Furthermore, the qualities of the ordering τ and of the one-bit p-values p_j both contribute to the power of knockoffs. Due to these challenges, no optimality results are currently available for knockoffs. We take a first step in this direction by exhibiting the test statistics T_j that lead to most powerful one-bit p-values against a point alternative.

Theorem 5. *Let $\bar{\mathcal{L}}$ be a fixed alternative distribution for $(\boldsymbol{X}, \boldsymbol{Y})$, with $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}) = \bar{f}(\boldsymbol{Y}|\boldsymbol{X})$. Define the likelihood statistic*

$$T_j^{\text{opt}}([X, \widetilde{X}], Y) \equiv \prod_{i=1}^{n} \bar{f}(Y_i | X_{i,\bullet}). \tag{53}$$

Assuming that ties do not occur, that is

$$\mathbb{P}_{\bar{\mathcal{L}}}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] = 0, \tag{54}$$

we have that the above likelihood statistic yields an optimal one-bit p-value:

$$T_j^{\text{opt}} \in \arg\max_{T_j} \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y)]. \tag{55}$$

This theorem is proved in Appendix D. The reader observes that T_j^{opt} is not a function of the knockoff variables or of the index j, which may at first seem paradoxical. Recall from the definition (52), however, that the one-bit p-value compares the test statistic on the original augmented design $[X, \widetilde{X}]$ and its swapped version $[X, \widetilde{X}]_{\text{swap}(j)}$. Therefore, the optimal one-bit p-value checks whether the original jth variable $X_{\bullet,j}$ fits with the rest of the data better than does its knockoff $\widetilde{X}_{\bullet,j}$. Therefore, the optimal one-bit p-value is in fact a function of the knockoffs as well as the index j. A simple way of operationalizing Theorem 5 is to fit a model $\widehat{f}(\boldsymbol{Y}|\boldsymbol{X})$ based on $([X, \widetilde{X}], Y)$ in any way that treats original variables and knockoffs symmetrically, and then defining $T_j([X, \widetilde{X}], Y) \equiv \widehat{f}(Y|X)$. The above result continues to hold when $\boldsymbol{X}_j$ is a *group* of variables, giving a clean way to combine evidence across multiple variables. A conditional version of the optimality statement (55) holds; see equation (112) in the appendix.

Theorem 5 requires that ties occur with probability zero (54). Proposition 1 below (proved in Appendix C) states that this nondegeneracy condition holds if either $\boldsymbol{Y}|\boldsymbol{X}$ or $\boldsymbol{X}_j|\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has a continuous distribution.

Proposition 1. *Suppose $\bar{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{X}) = g_{\boldsymbol{\eta}}$, where $\boldsymbol{\eta} = \boldsymbol{X}_j\beta_j + f_{-j}(\boldsymbol{X}_{-j})$ and g_η is a one-dimensional exponential family with natural parameter η and strictly convex, continuous log partition function ψ. Suppose also that $\boldsymbol{X}_j, \beta_j \in \mathbb{R}$, with $\beta_j \neq 0$. The nondegeneracy condition* (54) *holds if either*

1. *$\boldsymbol{X}_j|\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has a density for each $\boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$, or*
2. *g_η has a density,*

where the densities are with respect to the Lebesgue measure.

Finally, we remark that there are a few existing power analyses for knockoffs, all in high-dimensional asymptotic regimes and assuming lasso-based test statistics. Weinstein et al [**Weinstein2017**] analyze the power of a knockoffs variant in the case of independent Gaussian covariates, while Liu and Rigollet [**Liu2019**] and Fan et al [**Fan2020**] study conditions for consistency under correlated designs. Our finite-sample optimality result for the likelihood statistic is complementary to these previous works. Recently, Theorem 5 inspired a more powerful variant of knockoffs based on *masked likelihood ratio* statistics, which comes with a different kind of optimality guarantee [**Spector2022**].

6 Discussion

In this paper, we gave some answers to the theoretical questions posed in the introduction. We presented the first finite-sample optimality results in the MX framework and explicitly quantified how the performance of the underlying machine learning procedure impacts the asymptotic power of the CRT. Along the way, we exhibited a weakened form of the MX assumption and a resampling-free methodology valid under only this assumption.

The MX framework is just one setting where black-box prediction methods have been recently employed for the purpose of more powerful statistical inference. Other examples include conformal prediction [**FoygelBarber2019**], classification-based two-sample testing [**Kim2020**] and data-carving based multiple testing [**lei2016adapt**]. These methods employ machine learning algorithms to create powerful test statistics, calibrating them for valid inference with no assumptions about the method used. However, the more accurate the learned model, the more powerful the inference. Our finite-sample and asymptotic power results explicitly tie the error of the learning algorithm to the power of the test, and thus put this common intuition on a quantitative foundation and may thus help inform the choice and design of machine learning methods used for inferential goals.

Another set of connections we highlighted throughout the paper is to causal inference and semiparametric estimation. The MX CI problem has strong similarities to the problem of testing Fisher's strong null in a randomized experiment with potentially non-binary treatment and known propensity function. Furthermore, the CRT is similar in spirit to the Fisher randomization test. We believe these connections can be further leveraged to address problems in the MX framework that remain open. For example, consider the situation when the MX assumption is only approximately correct. This is analogous to the situation in observational studies, where the propensity score/function must be estimated. There is a vast literature on this topic based on "double robustness/machine-learning" [**Chernozhukov2018**] or targeted learning [**VanderLaan2011**]. Similar ideas may help relax the MX assumption [**Huang2019**] or study robustness to its misspecification [**Barber2018**]. Another topic that has received little attention in the MX community is that of estimation (with the exception of [**Zhang2020**]). Causal inference is a rich source of meaningful estimands (such as the *dose response function* [**Hirano2004**]) and estimators (such as the proposal of Kennedy et al. [**Kennedy2017**] for doubly-robust dose response function estimation). Such ideas may be directly relevant to the MX framework.

Much still remains to be done to systematically understand the theoretical properties of MX methods. One interesting direction is to analyze the case when $\widehat{g}_n$ is learned on the same data as is used for testing. We saw in Section 4.3 that Theorem 4 extends to lasso-based estimators $\widehat{g}_n$ learned in-sample, but the generality of such results remains an open question. It would also be interesting to consider alternatives beyond the linear model (17). A natural next step would be to consider generalized linear models. Furthermore, the connections to causal inference referenced above are tantalizing and deserve a dedicated treatment. Finally, we hope that these new theoretical insights about MX methods will lead to improved methodologies that are both statistically and computa-

tionally efficient, along the lines of the CRT variants discussed in this paper and in recent work [**Liu2020**].

Acknowledgments

We thank Asaf Weinstein, Timothy Barry, and Stephen Bates for detailed comments on earlier versions of the manuscript, as well as Ed Kennedy and Larry Wasserman for discussions of the connections to causal inference. We also thank two anonymous referees for constructive feedback that greatly helped us improve the manuscript.

A Proofs of Theorem 1

Proof. Let ϕ be any test satisfying conditional validity property (11). Let $\mathcal{A}$ be a set of pairs (y, z), for which both ϕ and $\phi^{\text{CRT}}_{T^{\text{opt}}}$ have level α conditionally on $Y = y, Z = z$. By assumption, $\mathbb{P}[(Y, Z) \in \mathcal{A}] = 1$. Now, fix realizations $(y, z) \in \mathcal{A}$. We first claim that the conditional power of ϕ is bounded above by that of $\phi^{\text{CRT}}_{T^{\text{opt}}}$, i.e.

$$\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X, y, z)|Y = y, Z = z] \leq \mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\text{CRT}}_{T^{\text{opt}}}(X, y, z)|Y = y, Z = z] \tag{56}$$

In the conditional problem, the alternative $\bar{\mathcal{L}}$ induces the following distribution for X:

$$\bar{\mathcal{L}}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i)\frac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)}, \tag{57}$$

where

$$\bar{f}(y_i|z_i) \equiv \int \bar{f}(y_i|x_i, z_i)f^*(x_i|z_i)dx_i.$$

The conditional problem is therefore a test of

$$H_0 : \mathcal{L}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i) \quad \text{versus}$$
$$H_1 : \mathcal{L}(X = x|Y = y, Z = z) = \prod_{i=1}^{n} f^*(x_i|z_i)\frac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)}.$$

This is a simple testing problem, with point null and point alternative. By the Neyman-Pearson lemma, the most powerful test is the one that rejects for large values of the likelihood ratio

$$\frac{\prod_{i=1}^{n} f^*(x_i|z_i)\frac{\bar{f}(y_i|x_i,z_i)}{\bar{f}(y_i|z_i)}}{\prod_{i=1}^{n} f^*(x_i|z_i)} = \prod_{i=1}^{n} \frac{\bar{f}(y_i|x_i, z_i)}{\bar{f}(y_i|z_i)} \propto T^{\text{opt}}(x, y, z), \tag{58}$$

verifying the conditional optimality claim (56). To obtain the unconditional claim (14), we take an expectation over Y, Z and use the fact that $\mathbb{P}[(Y,Z) \in \mathcal{A}] = 1$:

$$\begin{aligned}
\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,Y,Z)] &= \mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,Y,Z) \mid (Y,Z) \in \mathcal{A}] \\
&= \mathbb{E}_{\bar{\mathcal{L}}}[\mathbb{E}_{\bar{\mathcal{L}}}[\phi(X,Y,Z) \mid Y,Z] \mid (Y,Z) \in \mathcal{A}] \\
&\leq \mathbb{E}_{\bar{\mathcal{L}}}\left[\mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}(X,Y,Z) \mid Y,Z] \mid (Y,Z) \in \mathcal{A}\right] \\
&= \mathbb{E}_{\bar{\mathcal{L}}}\left[\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}(X,Y,Z) \mid (Y,Z) \in \mathcal{A}\right] = \mathbb{E}_{\bar{\mathcal{L}}}[\phi^{\mathrm{CRT}}_{T^{\mathrm{opt}}}(X,Y,Z)].
\end{aligned} \tag{59}$$

This completes the proof. □

B Simulation: Finite sample error control of the MX(2) F-test

In this section, we examine via numerical simulation the Type-I error control of the MX(2) F-test in finite samples, both if $\widehat{g}_n$ is fit out of sample (the case covered by Theorem 3) and if $\widehat{g}_n$ is fit in sample (conjectured). Code to reproduce the simulation is available online at `https://github.com/ekatsevi/crtpower-manuscript`.

Simulation setup. Recall that the MX(2) F-test is equivalent to the dCRT in finite samples when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is Gaussian. Therefore, to test the Type-I error control of the MX(2) F-test in a nontrivial setting, we instead consider a discrete distribution for $(\boldsymbol{X}, \boldsymbol{Z})$. In particular, we sample $(\boldsymbol{X}, \boldsymbol{Z})$ from a Markov chain, as described next. (Such a Markovian setup has often been employed in MX analyses of GWAS studies [**SetC17**, **SetS19**, **Bates2020**], motivated by recombination models from population genetics.)

Let's assume for simplicity that $\dim(\boldsymbol{X}) = 1$. Define $(\boldsymbol{X}, \boldsymbol{Z}) \in \{0,1\}^{1+p}$ to have the distribution of a Markov chain with

$$\text{initial state } \boldsymbol{X} \sim \mathrm{Ber}(\pi_{\mathrm{init}}) \text{ and transition matrix } \begin{pmatrix} 1-\pi_{\mathrm{flip}} & \pi_{\mathrm{flip}} \\ \pi_{\mathrm{flip}} & 1-\pi_{\mathrm{flip}} \end{pmatrix}.$$

More explicitly, we have

$$\mathbb{P}[\boldsymbol{X} = x, \boldsymbol{Z} = z] = \pi_{\mathrm{init}}^{x}(1-\pi_{\mathrm{init}})^{1-x}\pi_{\mathrm{flip}}^{\mathbb{1}(z_1 \neq x)}(1-\pi_{\mathrm{flip}})^{\mathbb{1}(z_1 = x)} \prod_{j=2}^{p} \pi_{\mathrm{flip}}^{\mathbb{1}(z_j \neq z_{j-1})}(1-\pi_{\mathrm{flip}})^{\mathbb{1}(z_j = z_{j-1})}.$$

The parameters $(\pi_{\mathrm{init}}, \pi_{\mathrm{flip}})$ describe the distribution of $\boldsymbol{X}|\boldsymbol{Z}$ and are assumed known. Furthermore, let the response $\boldsymbol{Y}$ be distributed as a random effects model in $\boldsymbol{Z}$:

$$\boldsymbol{Y} = \boldsymbol{Z}^T\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \quad \boldsymbol{\gamma} \sim N(0, \sigma_\gamma^2 I_p),\ \boldsymbol{\epsilon} \sim N(0, \sigma_\epsilon^2 I_n).$$

Thus, all simulations are conducted under the null hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. The signal-to-noise ratio in this relationship is defined via

$$\mathrm{SNR} = \frac{\mathbb{E}[\|\boldsymbol{Z}\|^2]\sigma_\gamma^2}{\sigma_\epsilon^2}.$$

Suppose we have n_{train} and n_{test} training and test samples, respectively. Then, the function $\widehat{g}_n$ is defined by running a 10-fold cross-validated ridge regression of Y on Z using either the n_{train} training samples (out of sample training) or the n_{test} test samples (in sample training) and then the statistic $U_n(X, Y, Z)$ is computed using the n_{test} test samples.

Simulation parameters. All simulations were run with

$$n_{\text{train}} = 100; \quad \pi_{\text{init}} = 0.1; \quad \pi_{\text{flip}} = 0.1; \quad \sigma_\epsilon^2 = 1. \tag{60}$$

On the other hand, the three parameters $(n_{\text{test}}, \text{SNR}, p)$ were varied as follows:

$$n_{\text{test}} \in \{10, 25, \mathbf{100}\}; \quad \text{SNR} \in \{0, \mathbf{1}, 5\}; \quad p \in \{20, 100, \mathbf{500}\}.$$

The bolded values above represent the *default values* for each parameter. Each of the three parameters was varied while keeping the other two parameters at their default values, giving a total of nine simulation settings. For each simulation setting, the training data were generated just once, since Theorem 3 implicitly conditions on the training data. The entire test data (X, Y, Z) were sampled 1000 times to generate the null distribution of $U_n(X, Y, Z)$.

Simulation results. For each of the nine simulation settings, we produce normal QQ plots of the z-statistics $U_n(X, Y, Z)$ based on out of sample or in sample training (Figures 1 and 2, respectively). When $\widehat{g}_n$ is fit out of sample (Figure 1), we see good calibration in most cases. In particular, the test sample size impacts calibration, but the SNR and the dimension do not. The test statistic's null distribution shows some inflation for the small sample size of $n_{\text{test}} = 10$, but is already well-calibrated starting with $n_{\text{test}} = 25$. Therefore, the asymptotic Type-I error control proved in Theorem 3 extends to modest sample sizes as well. Furthermore, when $\widehat{g}_n$ is fit in sample (Figure 2), we observe calibration that is as good as when $\widehat{g}_n$ is fit out of sample. This suggests that we may apply the MX(2) F-test even with in-sample-estimated $\widehat{g}_n$. We must bear in mind, however, that different choices of the fixed parameters (60) may alter these conclusions. In particular, smaller π_{init} leads to more discreteness in X and therefore slower convergence to normality.

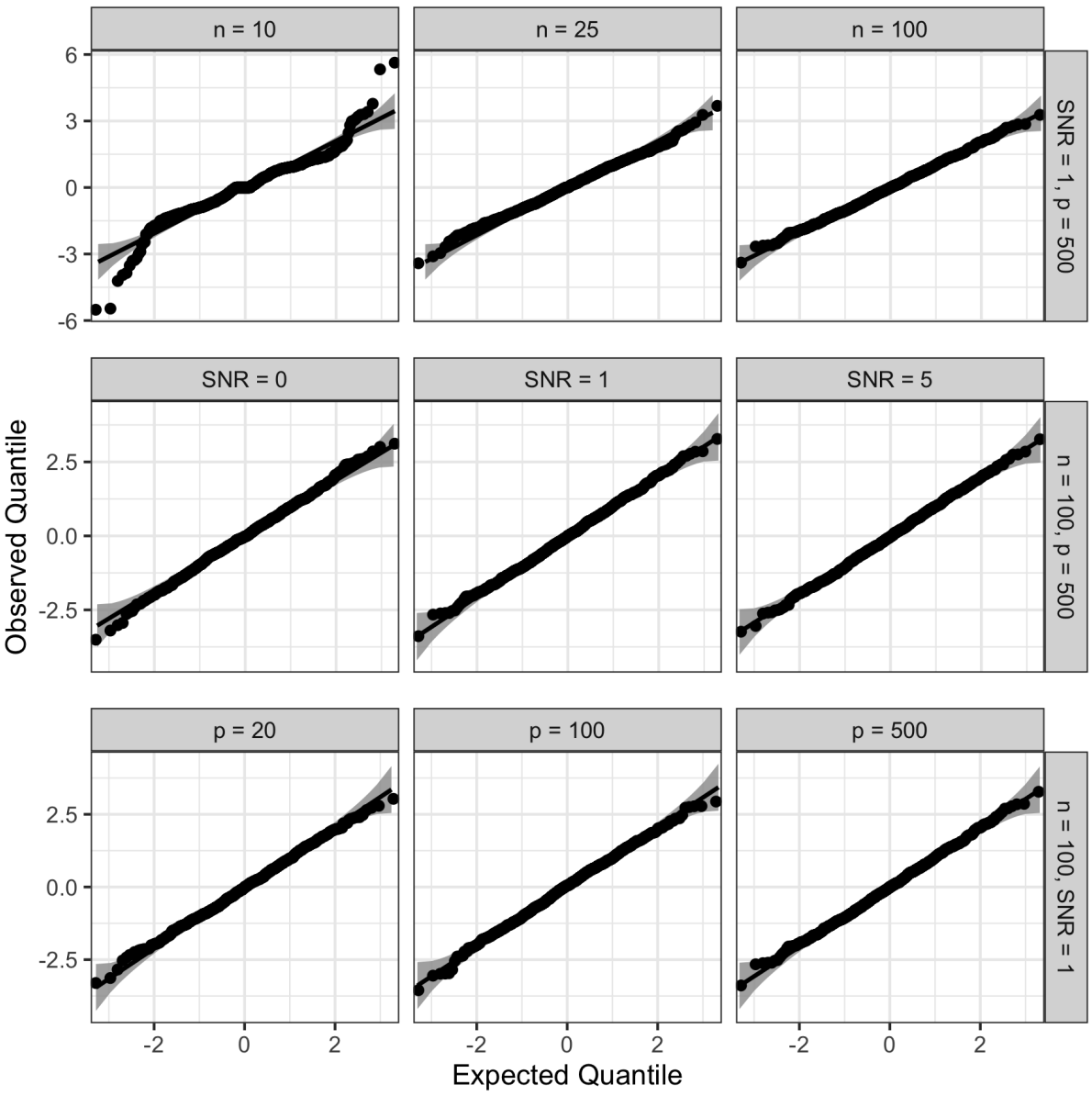

Figure 1: Distributions of 1000 samples of $U_n(X, Y, Z)$ each from nine simulation settings under the null, where $\widehat{g}_n$ is learned *out of sample*.

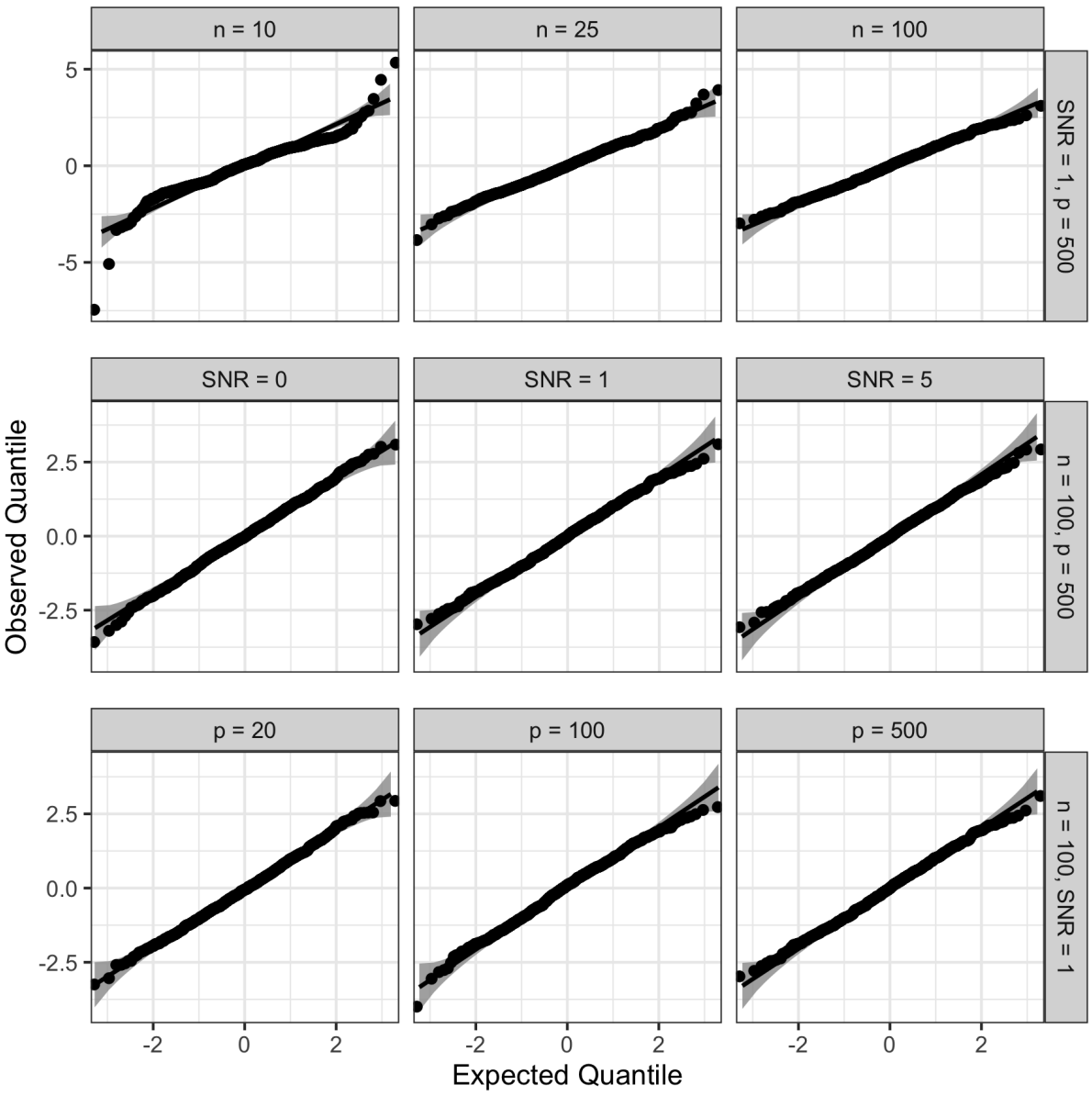

Figure 2: Distributions of 1000 samples of $U_n(X, Y, Z)$ each from nine simulation settings under the null, where $\widehat{g}_n$ is learned *in sample*.

C Proofs for Sections 3 and 4

C.1 Proofs of main results

Proof of Theorem 2. First, conclusion (90) of Lemma 3—which applies because of the assumption $\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$—states that for

$$\widetilde{X}_i^1, \widetilde{X}_i^2 | Y, Z \overset{\text{ind}}{\sim} \mathcal{L}_n(\boldsymbol{X} | \boldsymbol{Z} = Z_i),$$

we have the convergence

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z) \end{pmatrix} \overset{\mathcal{L}_n}{\to}_d N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix} \right). \tag{61}$$

By the continuous mapping theorem, we find that

$$(T_n(\widetilde{X}^1, Y, Z), T_n(\widetilde{X}^2, Y, Z)) \overset{\mathcal{L}_n}{\to}_d \chi_d^2 \times \chi_d^2. \tag{62}$$

Since χ_d^2 has a continuous and strictly increasing distribution function, we conclude using Lemma 1 that $C_n(Y, Z) \overset{\mathcal{L}_n}{\to}_p Q_{1-\alpha}[\chi_d^2] = c_{d,1-\alpha}$, proving the statement (31).

Next, note that for any $\delta > 0$,

$$\begin{aligned}
&\mathbb{P}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z) \neq \phi_n^{\mathcal{L}_n}(X, Y, Z)] \\
&\quad = \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z))] \\
&\quad = \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z)), |C_n(Y, Z) - c_{d,1-\alpha}| \leq \delta] \\
&\qquad + \mathbb{P}_{\mathcal{L}_n}[\min(c_{d,1-\alpha}, C_n(Y, Z)) < T_n(X, Y, Z) \leq \max(c_{d,1-\alpha}, C_n(Y, Z)), |C_n(Y, Z) - c_{d,1-\alpha}| > \delta] \\
&\quad \leq \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] + \mathbb{P}_{\mathcal{L}_n}[|C_n(Y, Z) - c_{d,1-\alpha}| > \delta].
\end{aligned}$$

To justify the last step, suppose without loss of generality that $c_{d,1-\alpha} \leq C_n(Y, Z)$. Then, note that if $c_{d,1-\alpha} < T_n(X, Y, Z) \leq C_n(Y, Z)$ and $C_n(Y, Z) - c_{d,1-\alpha} \leq \delta$ then

$$|T_n(X, Y, Z) - c_{d,1-\alpha}| = T_n(X, Y, Z) - c_{d,1-\alpha} \leq C_n(Y, Z) - c_{d,1-\alpha} \leq \delta.$$

Taking a lim sup on both sides in the display before the last and using the convergence $C_n(Y, Z) \overset{\mathcal{L}_n}{\to}_p c_{d,1-\alpha}$, we find that

$$\begin{aligned}
&\limsup_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z) \neq \phi_n^{\mathcal{L}_n}(X, Y, Z)] \\
&\quad \leq \limsup_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] + \limsup_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[|C_n(Y, Z) - c_{d,1-\alpha}| > \delta] \\
&\quad = \limsup_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta].
\end{aligned}$$

Letting $\delta \to 0$ and using the assumption (32), we arrive at the claimed asymptotic equivalence (33). This completes the proof. □

Proof of Theorem 3. Fix any sequence $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$. Because $\mathcal{L}_n \in \mathscr{L}_0$, we have $(X, Y, Z) \overset{d}{=} (\widetilde{X}, Y, Z)$, where $\widetilde{X}_i | Y, Z \overset{\text{ind}}{\sim} \mathcal{L}_n(\boldsymbol{X} | \boldsymbol{Z} = Z_i)$. By conclusion (90) of Lemma 3, which applies because $\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$ by assumption, we have $U_n(X, Y, Z) \overset{d}{=} U_n(\widetilde{X}, Y, Z) \overset{\mathcal{L}_n}{\to}_d N(0, I_d)$. This verifies the asymptotic normality statement (36).

To show the asymptotic Type-I error control statement (37), it suffices to show that for any sequence $\mathcal{L}_n \in \mathscr{L}_0^{\mathrm{MX}(2)} \cap \mathscr{L}_n(c_1, c_2)$, we have

$$\limsup_{n \to \infty} \ \mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z)] \leq \alpha. \tag{63}$$

By the continuous mapping theorem it follows from asymptotic normality (36) that $T_n(X, Y, Z) = \|U_n(X, Y, Z)\|^2 \overset{\mathcal{L}_n}{\to}_d \chi_d^2$. Therefore,

$$\lim_{n \to \infty} \ \mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z)] = \lim_{n \to \infty} \ \mathbb{P}_{\mathcal{L}_n}[T_n(X, Y, Z) > c_{d,1-\alpha}] = \mathbb{P}[\chi_d^2 > c_{d,1-\alpha}] = \alpha,$$

from which the conclusion (63) follows. This completes the proof. □

Proof of Theorem 4. In Lemma 4, we show that the estimator $\widehat{\rho}_n$ is consistent, i.e.

$$\widehat{\rho}_n \overset{\mathcal{L}_n}{\to}_p \overline{\Sigma}\beta. \tag{64}$$

Next, we derive that

$$T_n(X, Y, Z) = \|\sqrt{n}\widehat{S}_n^{-1}\widehat{\rho}_n\|^2 = \|\sqrt{n}\widehat{S}_n^{-1}S_n S_n^{-1}\widehat{\rho}_n\|^2 \geq \left(\sqrt{n}\lambda_{\min}(\widehat{S}_n^{-1}S_n)\lambda_{\min}(S_n^{-1})\|\widehat{\rho}_n\|\right)^2.$$

Now, we have $\widehat{S}_n^{-1}S_n \overset{\mathcal{L}_n}{\to}_p I_d$ by conclusion (83) of Lemma 2, so the continuous mapping theorem implies that $\lambda_{\min}(\widehat{S}_n^{-1}S_n) \overset{\mathcal{L}_n}{\to}_p 1$. Furthermore, $\inf_n \lambda_{\min}(S_n^{-1}) > 0$ by conclusion (102) of Lemma 5. Finally, $\|\widehat{\rho}_n\| \overset{\mathcal{L}_n}{\to}_p \|\overline{\Sigma}\beta\|$ by equation (64), and

$$\|\overline{\Sigma}\beta\| \geq \lambda_{\min}(\overline{\Sigma})\|\beta\| = \|\overline{\Sigma}^{-1}\|^{-1}\|\beta\| \geq \left(\sup_n \|\overline{\Sigma}_n^{-1}\|\right)^{-1} \|\beta\| > 0,$$

since $\beta \neq 0$ by assumption and assumptions (40) and (43) imply that $\|\overline{\Sigma}^{-1}\| \leq \sup_n \|\overline{\Sigma}_n^{-1}\| < \infty$. Putting these facts together implies that $\sqrt{n}\lambda_{\min}(\widehat{S}_n^{-1}S_n)\lambda_{\min}(S_n^{-1})\|\widehat{\rho}_n\| \overset{\mathcal{L}_n}{\to}_p \infty$, and therefore $T_n(X, Y, Z) \overset{\mathcal{L}_n}{\to}_p \infty$. Hence,

$$\mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z)] = \mathbb{P}_{\mathcal{L}_n}[T_n(X, Y, Z) > c_{d,1-\alpha}] \to 1. \tag{65}$$

The fact that $T_n(X, Y, Z) \overset{\mathcal{L}_n}{\to}_p \infty$ also implies that $\limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = 0$ for any $\delta > 0$. Hence, the condition (32) of Theorem 2 is satisfied, so the conclusion (33) implies that

$$\lim_{n \to \infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\mathcal{L}_n}(X, Y, Z)] = \lim_{n \to \infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{N(\mu_n, \Sigma_n)}(X, Y, Z)] = 1.$$

Thus, we have shown the claimed consistency (44), so we have finished the proof of part (a) of the theorem.

To prove part (b), we claim that it suffices to establish that

$$T_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d \chi^2_d(\|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2}h\|^2). \tag{66}$$

Indeed, the limiting power of the MX(2) F-test would directly follow from this statement. To establish that the CRT has the same limiting power, by Theorem 2 it suffices to verify the non-accumulation condition (32). Letting T be the limiting distribution in claim (66), this claim implies that for any $\delta > 0$,

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - c_{d,1-\alpha}| \leq \delta] = \mathbb{P}[|T - c_{d,1-\alpha}| \leq \delta].$$

Because T has a continuous distribution function, the limit above tends to zero. Therefore, it is indeed sufficient to verify the claimed convergence (66). This statement, in turn, will follow if we prove that

$$U_n(X, Y, Z) \xrightarrow{\mathcal{L}_n}_d N((\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma}h, I_d). \tag{67}$$

Indeed, note that

$$h^T\overline{\Sigma}(\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1}\overline{\Sigma}h = h^T\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)^{-1}\overline{\Sigma}^{1/2}h = \|(\sigma^2 I_d + \mathcal{E}^2)^{-1/2}\overline{\Sigma}^{1/2}h\|^2.$$

To show the statement (67), we first rewrite $U_n(X, Y, Z)$ as follows:

$$\begin{aligned}
&U_n(X, Y, Z) \\
&= \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n ((X_i - \mu_n(Z_i))^T\beta_n + \epsilon_i + \bar{g}_n(Z_i) - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \\
&= \frac{\widehat{S}_n^{-1}}{n} \sum_{i=1}^n (X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T h_n + \frac{\widehat{S}_n^{-1}}{\sqrt{n}} \sum_{i=1}^n (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \\
&\equiv A_n + B_n,
\end{aligned}$$

where $Y_i' \equiv \bar{g}_n(Z_i) + \epsilon_i$. It therefore suffices to show that

$$A_n \xrightarrow{\mathcal{L}_n}_p (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma}h \quad \text{and} \quad B_n \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{68}$$

By conclusion (101) of Lemma 5, there exist c_1, c_2 for which $\mathcal{L}_n \in \mathscr{L}(c_1, c_2)$ for each n. Therefore, we can apply Lemma 2 to conclude that

$$\widehat{S}_n^{-1} S_n \xrightarrow{\mathcal{L}_n}_p I_d. \tag{69}$$

By conclusion (103) of Lemma 5, we have that $S_n^2 \to \overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2}$, so

$$S_n^{-1} \to (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}. \tag{70}$$

Now, we apply the WLLN to find the limit of A_n. Since $(X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T$ has expectation $\overline{\Sigma}$ and second moment uniformly bounded by the eighth moment assumption (41), we can apply the weak law of large numbers as well as the statements (69) and (70) to conclude that

$$A_n = (\widehat{S}_n^{-1} S_n) \frac{S_n^{-1}}{n} \sum_{i=1}^{n} (X_i - \mu_n(Z_i))(X_i - \mu_n(Z_i))^T h_n \xrightarrow{\mathcal{L}_n}_p (\overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2})^{-1/2}\overline{\Sigma} h.$$

Next, we seek to find the limit of B_n. Defining $\boldsymbol{Y}', S_n'^2, \mathcal{L}_n'$ according to (100) below, we may rewrite

$$B_n = (\widehat{S}_n^{-1} S_n)(S_n^{-1} S_n') \frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)). \tag{71}$$

By conclusion (103) of Lemma 5, we have

$$S_n^{-1} S_n' \to I_d. \tag{72}$$

Furthermore, conclusion (101) of Lemma 5 gives $\mathcal{L}_n' \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2)$. Therefore, $\mathcal{L}_n'$ satisfies the assumptions of Lemma 3, statement (89) of which gives

$$\frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(\widetilde{X}_i - \mu_n(Z_i)) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{73}$$

Furthermore, $\mathcal{L}_n' \in \mathscr{L}_0$ implies that $(\boldsymbol{X}, \boldsymbol{Y}', \boldsymbol{Z}) \overset{d}{=} (\widetilde{\boldsymbol{X}}, \boldsymbol{Y}', \boldsymbol{Z})$, which together with the convergence (73) implies that

$$\frac{S_n'^{-1}}{\sqrt{n}} \sum_{i=1}^{n} (Y_i' - \widehat{g}_n(Z_i))(X_i - \mu_n(Z_i)) \xrightarrow{\mathcal{L}_n}_d N(0, I_d). \tag{74}$$

Finally, putting together displays (69), (72) and (74) yields that $B_n \xrightarrow{\mathcal{L}_n}_d N(0, I_d)$. This verifies the claimed convergences (68) and therefore completes the proof. □

Proof of Corollary 2. First we verify the statement (47). To this end, first note that

$$\begin{aligned}
\mathcal{E}_n^2 &= \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - \bar{g}_n(\boldsymbol{Z}))^2 \overline{\Sigma}_n^{-1/2} \Sigma_n(\boldsymbol{Z}) \overline{\Sigma}_n^{-1/2} \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - g_n(\boldsymbol{Z}))^2 \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\widehat{\gamma}_{\pi n} - \gamma_n)^T \boldsymbol{Z}\boldsymbol{Z}^T (\widehat{\gamma}_{\pi n} - \gamma_n) \mid X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&= (\widehat{\gamma}_{\pi n} - \gamma_n)^T \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Z}\boldsymbol{Z}^T](\widehat{\gamma}_{\pi n} - \gamma_n) \\
&= \|\widehat{\gamma}_{\pi n} - \gamma_n\|^2.
\end{aligned} \tag{75}$$

The second equality holds because for $(\boldsymbol{X}, \boldsymbol{Z})$ jointly Gaussian, $\Sigma_n(\boldsymbol{Z})$ is constant in $\boldsymbol{Z}$, so $\overline{\Sigma}_n^{-1/2} \Sigma_n(\boldsymbol{Z}) \overline{\Sigma}_n^{-1/2} = 1$. Therefore, the variance-weighted mean-squared error $\mathcal{E}_n^2$ of

$\widehat{g}_n$ reduces to the squared error in the estimate $\widehat{\gamma}_{\pi n}$. To obtain the limit of the latter quantity, we appeal to Bayati and Montanari's Corollary 1.6 [**Bayati2011**]. To verify the conditions of this corollary, it suffices to verify part (b) of their Definition 1: that the empirical distribution of the noise terms $\epsilon_i' \equiv Y_i - Z_i^T \gamma_n = X_i \beta_n + \epsilon_i$ in the training set (say $1 \leq i \leq \pi n$) converges weakly to a random variable Λ and $\frac{1}{\pi n} \sum_{i=1}^{\pi n} \epsilon_i'^2 \to \mathbb{E}[\Lambda^2]$. These statements hold almost surely in the training data by the strong law of large numbers if we assume without loss of generality that X_i and ϵ_i are both defined as the first πn elements of infinite i.i.d. sequences with distributions $N(0,1)$ and $N(0,\sigma^2)$, respectively. Therefore, Bayati and Montanari's Corollary 1.6 gives

$$\frac{1}{p} \| \sqrt{\pi n} \widehat{\gamma}_{\pi n} - \sqrt{\pi n} \gamma_n \|^2 \overset{\text{a.s.}}{\to} \pi \delta (\tau_*^2 - \sigma^2), \tag{76}$$

where the almost sure statement is with respect to the training data $(X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}})$. Since $\pi n / p \to \pi \delta$, we can cancel these terms from the above equation to obtain

$$\| \widehat{\gamma}_{\pi n} - \gamma_n \|^2 \overset{\text{a.s.}}{\to} \tau_*^2 - \sigma^2. \tag{77}$$

Putting together equations (75) and (77) gives the claimed statement (47).

To apply this result, we must verify the assumptions of Theorem 4. The bounded inverse assumption (40) holds because $\overline{\Sigma}_n = 1$ for all n in the orthogonal design setting. The eighth moment assumption (41) holds due to the boundedness of the eighth moments of Gaussian random variables. To verify the moment assumption (42), we note that, almost surely in the training data,

$$\begin{aligned}
&\sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\widehat{g}_n(\boldsymbol{Z}) - \bar{g}_n(\boldsymbol{Z}))^4 \| \boldsymbol{X} - \mu_n(\boldsymbol{Z}) \|^4 | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&\quad = \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Z} \widehat{\gamma}_{\pi n} - \boldsymbol{Z} \gamma_n)^4 \| \boldsymbol{X} \|^4 | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \\
&\quad \leq \sup_n \| \widehat{\gamma}_{\pi n} - \gamma_n \|^4 \mathbb{E}_{\mathcal{L}_n}[\| \boldsymbol{Z} \|^4 \| \boldsymbol{X} \|^4] \\
&\quad \leq \sup_n \| \widehat{\gamma}_{\pi n} - \gamma_n \|^4 \mathbb{E}_{\mathcal{L}_n}[\| \boldsymbol{Z} \|^8]^{1/2} \mathbb{E}_{\mathcal{L}_n}[\| \boldsymbol{X} \|^8]^{1/2} \\
&\quad < \infty.
\end{aligned}$$

The last inequality holds because $\| \widehat{\gamma}_{\pi n} - \gamma_n \|^4$ has a finite limit according to (77) and because $\boldsymbol{Z}$ and $\boldsymbol{X}$ have bounded eighth moments since they are Gaussian. Finally, we verify assumption (43) by noting that $\overline{\Sigma}_n \to \overline{\Sigma} \equiv 1$ and $\mathcal{E}_n^2 \to \tau_*^2 - \sigma^2 \equiv \mathcal{E}$, the latter by statement (47). Therefore, Theorem 4 gives

$$\mathbb{E}_{\mathcal{L}_n}[\phi^{\mathcal{L}_n}_{(1-\pi)n}(X_{\text{test}}, Y_{\text{test}}, Z_{\text{test}}) | X_{\text{train}}, Y_{\text{train}}, Z_{\text{train}}] \overset{\text{a.s.}}{\to} \mathbb{P}[\chi_1^2(\| \tau_*^{-1} h \sqrt{1-\pi} \|^2) > c_{1,1-\alpha}].$$

The extra factor of $\sqrt{1-\pi}$ reflects the fact that a sample size of $(1-\pi)n$ is used for testing, so $\beta_n = h_n / \sqrt{n} = h_n \sqrt{1-\pi} / \sqrt{(1-\pi)n}$. In other words, reducing the number of samples for testing from n to $(1-\pi)n$ has the effect of reducing the alternative signal strength

from h_n to $h_n\sqrt{1-\pi}$. Noting that $c_{1,1-\alpha} = z^2_{1-\alpha/2}$, we conclude using the dominated convergence theorem that

$$\begin{aligned}\mathbb{E}_{\mathcal{L}_n}[\varphi_n^{\mathcal{L}_n}(X,Y,Z)] &= \mathbb{E}_{\mathcal{L}_n}\left[\mathbb{E}_{\mathcal{L}_n}[\phi^{\mathcal{L}_n}_{(1-\pi)n}(X_{\text{test}},Y_{\text{test}},Z_{\text{test}})|X_{\text{train}},Y_{\text{train}},Z_{\text{train}}]\right] \\ &\to \mathbb{P}[|N(\tau_*^{-1}h\sqrt{1-\pi},1)| > z_{1-\alpha/2}].\end{aligned}$$

This completes the proof of the corollary. □

C.2 Technical lemmas

First, we state a lemma that gives a sufficient condition for the convergence of the CRT threshold, which follows directly from Lemmas 2 and 3 of [**Wang2020b**].

Lemma 1 ([**Wang2020b**])**.** *Let $\mathcal{L}_n$ be a sequence of laws over $(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$, from which (X,Y,Z) are sampled. Furthermore, for each i sample two independent copies*

$$\widetilde{X}_i^1, \widetilde{X}_i^2 \overset{i.i.d.}{\sim} \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}=Z_i) \quad \textit{such that, given } Z, (\widetilde{X}_1^1,\widetilde{X}_1^2) \perp\!\!\!\perp \cdots \perp\!\!\!\perp (\widetilde{X}_n^1,\widetilde{X}_n^2) \perp\!\!\!\perp Y. \tag{78}$$

Suppose that $T_n(X,Y,Z)$ is a test statistic satisfying

$$(T_n(\widetilde{X}^1,Y,Z), T_n(\widetilde{X}^2,Y,Z)) \overset{\mathcal{L}_n}{\to}_d \widetilde{T}\times\widetilde{T} \tag{79}$$

for some limiting random variable $\widetilde{T}$ with continuous and strictly increasing distribution function. Then, the CRT threshold converges in probability to the upper quantile of $\widetilde{T}$:

$$C_n(Y,Z) \equiv Q_{1-\alpha}[T_n(\widetilde{X},Y,Z)|Y,Z] \overset{\mathcal{L}_n}{\to}_p Q_{1-\alpha}[\widetilde{T}]. \tag{80}$$

Lemma 2. *Fix any $c_1, c_2 > 0$. For any sequence*

$$\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot),\Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1,c_2), \tag{81}$$

we have

$$\widehat{S}_n^2 - S_n^2 \overset{\mathcal{L}_n}{\to}_p 0 \tag{82}$$

and

$$\widehat{S}_n^{-1} S_n \overset{\mathcal{L}_n}{\to}_p I_d. \tag{83}$$

Proof. To show the first convergence (82), we apply the WLLN to the triangular array $\{(Y_i - \widehat{g}_n(Z_i))^2\Sigma_n(Z_i)\}_{i,n}$. We first verify the second moment condition:

$$\begin{aligned}&\sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})\|^2] \\ &\quad= \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\|\Sigma_n(\boldsymbol{Z})\|^2] \\ &\quad\le \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^2|\boldsymbol{Z}]^2] \\ &\quad\le \sup_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}-\widehat{g}_n(\boldsymbol{Z}))^4\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]] \\ &\quad\le c_2 < \infty.\end{aligned} \tag{84}$$

Therefore, by the WLLN we obtain the convergence

$$\widehat{S}_n^2 - S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))^2\Sigma_n(Z_i) - \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})] \xrightarrow{\mathcal{L}_n}_p 0. \tag{85}$$

To show the second convergence (83), note first that

$$\begin{aligned}\sup_n \|S_n^2\| &= \sup_n \|\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})]\| \\ &\leq \sup_n \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})\|^2]^{1/2} \leq c_2^{1/2},\end{aligned} \tag{86}$$

the last step having been derived in equation (84). Therefore, for every n, we have

$$S_n^2 \in \mathcal{S} \equiv \{S^2 : \|S^{-1}\| \leq c_1, \|S^2\| \leq c_2^{1/2}\}. \tag{87}$$

Since $\mathcal{S}$ is a compact subset of the open set of positive definite matrices, there exists a $\delta > 0$ such that $\mathcal{S}_\delta = \{S^2 : \|S^2 - S_0^2\| \leq \delta \text{ for some } S_0^2 \in \mathcal{S}\}$ is also a compact subset of the set of positive definite matrices. Since the function $S^2 \mapsto S^{-1}$ is continuous on the compact set $\mathcal{S}_\delta$, it must be uniformly continuous on this set as well. Fix $\gamma > 0$. By uniform continuity, there exists an $\eta > 0$ such that $\|S_1^2 - S_2^2\| \leq \eta$ implies that $\|S_1^{-1} - S_2^{-1}\| \leq \gamma$ for all $S_1^2, S_2^2 \in \mathcal{S}_\delta$. We therefore have that

$$\begin{aligned}\mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma] &= \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma, \widehat{S}_n^2 \in \mathcal{S}_\delta] + \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma, \widehat{S}_n^2 \notin \mathcal{S}_\delta] \\ &\leq \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \eta] + \mathbb{P}_{\mathcal{L}_n}[\widehat{S}_n^2 \notin \mathcal{S}_\delta] \\ &\leq \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \eta] + \mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^2 - S_n^2\| > \delta].\end{aligned}$$

Using the convergence (82), we find that the last expression tends to zero as $n \to \infty$, from which it follows that $\mathbb{P}_{\mathcal{L}_n}[\|\widehat{S}_n^{-1} - S_n^{-1}\| > \gamma] \to 0$ as $n \to \infty$. Therefore,

$$\widehat{S}_n^{-1} - S_n^{-1} \xrightarrow{\mathcal{L}_n}_p 0.$$

Multiplying this relation on the right by the bounded quantity S_n, we arrive at the statement (83), which concludes the proof. □

Lemma 3. *Consider generating* $(\widetilde{X}^1, \widetilde{X}^2, Y, Z)$ *according to* (78) *for a sequence of laws*

$$\mathcal{L}_n \in \mathscr{L}^{\mathrm{MX}(2)}(\mu_n(\cdot), \Sigma_n(\cdot)) \cap \mathscr{L}_n(c_1, c_2). \tag{88}$$

We have

$$n^{-1/2}\begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1}\end{pmatrix}\sum_{i=1}^{n}(Y_i - \widehat{g}_n(Z_i))\begin{pmatrix}\widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i)\end{pmatrix} \xrightarrow{\mathcal{L}_n}_d N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d\end{pmatrix}\right) \tag{89}$$

and

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z)\end{pmatrix} \xrightarrow{\mathcal{L}_n}_d N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix} I_d & 0 \\ 0 & I_d\end{pmatrix}\right). \tag{90}$$

Proof of Lemma 3. Note that

$$\begin{pmatrix} U_n(\widetilde{X}^1, Y, Z) \\ U_n(\widetilde{X}^2, Y, Z) \end{pmatrix} = \begin{pmatrix} \widehat{S}_n^{-1} S_n & 0 \\ 0 & \widehat{S}_n^{-1} S_n \end{pmatrix} \cdot n^{-1/2} \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \sum_{i=1}^{n} (Y_i - \widehat{g}_n(Z_i)) \begin{pmatrix} \widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i) \end{pmatrix}.$$

By Lemma 2, we have that $\widehat{S}_n^{-1} S_n \xrightarrow{\mathcal{L}_n}_p I_d$, so by Slutsky we find that the second statement (90) follows from the first (89). Therefore, it suffices to prove the latter convergence. To this end, we apply the CLT to the triangular array of vectors

$$\left\{ (Y_i - \widehat{g}_n(Z_i)) \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{X}_i^1 - \mu_n(Z_i) \\ \widetilde{X}_i^2 - \mu_n(Z_i) \end{pmatrix} \right\}_{i,n}. \tag{91}$$

To apply the CLT, we first verify the Lyapunov condition with $\delta = 1$:

$$\begin{aligned}
&\sup_n \ \mathbb{E}_{\mathcal{L}_n} \left[\left\| (\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})) \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2 - \mu_n(\boldsymbol{Z}) \end{pmatrix} \right\|^3 \right] \\
&\quad \le \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[|\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})|^3 (\|\widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z})\|^2 + \|\widetilde{\boldsymbol{X}}^2 - \mu_n(\boldsymbol{Z})\|^2)^{3/2} \right] \\
&\quad \le \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[|\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})|^3 C \left(\|\widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z})\|^3 + \|\widetilde{\boldsymbol{X}}^2 - \mu_n(\boldsymbol{Z})\|^3 \right) \right] \\
&\quad = 2C \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[|\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})|^3 \|\widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z})\|^3 \right] \\
&\quad \le 2C \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \|\widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z})\|^4 \right]^{3/4} \\
&\quad = 2C \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n} [\|\widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}] \right]^{3/4} \\
&\quad = 2C \sup_n \ \|S_n^{-1}\|^3 \mathbb{E}_{\mathcal{L}_n} \left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n} [\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}] \right]^{3/4} \\
&\quad \le 2C c_1^3 c_2^{3/4} < \infty.
\end{aligned} \tag{92}$$

Here C is chosen such that $(a+b)^{3/2} \le C(a^{3/2} + b^{3/2})$ for all $a, b \ge 0$. Next, it is easy to verify that

$$\mathbb{E}_{\mathcal{L}_n} \left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})) \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2 - \mu_n(\boldsymbol{Z}) \end{pmatrix} \right] = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{93}$$

and

$$\mathrm{Var}_{\mathcal{L}_n} \left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z})) \begin{pmatrix} S_n^{-1} & 0 \\ 0 & S_n^{-1} \end{pmatrix} \begin{pmatrix} \widetilde{\boldsymbol{X}}^1 - \mu_n(\boldsymbol{Z}) \\ \widetilde{\boldsymbol{X}}^2 - \mu_n(\boldsymbol{Z}) \end{pmatrix} \right] = \begin{pmatrix} I_d & 0 \\ 0 & I_d \end{pmatrix}. \tag{94}$$

By the CLT, the convergence (89) now follows. □

Lemma 4. *In the setting of Theorem 4(a), define*

$$\rho_n \equiv \mathbb{E}_{\mathcal{L}_n}[\mathrm{Cov}_{\mathcal{L}_n}[\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{Z}]] = \overline{\Sigma}_n \beta \quad \textit{and} \quad \rho \equiv \lim_{n \to \infty} \rho_n = \overline{\Sigma} \beta.$$

Under the assumptions of Theorem 4, the estimator $\widehat{\rho}_n$ defined in (21) *is consistent for ρ:*

$$\widehat{\rho}_n \xrightarrow{\mathcal{L}_n}_p \rho = \overline{\Sigma}\beta. \tag{95}$$

Proof. Note that $\widehat{\rho}_n$ is the mean of i.i.d. terms with expectation

$$\begin{aligned} &\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))] \\ &\quad = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta + \boldsymbol{\epsilon} + \bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))] = \overline{\Sigma}_n\beta. \end{aligned} \tag{96}$$

These terms also have bounded second moment, since

$$\begin{aligned} &\mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))\|^2] \\ &\quad = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T\beta + \boldsymbol{\epsilon} + \bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\ &\quad \le C\mathbb{E}_{\mathcal{L}_n}[(\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2\|\beta\|^2 + \boldsymbol{\epsilon}^2 + (\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2)\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\ &\quad = C\|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] + C\sigma^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2] \\ &\qquad + C\mathbb{E}_{\mathcal{L}_n}[(\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^2]. \end{aligned} \tag{97}$$

Here, C is a constant so that $(a+b+c)^2 \le C(a^2+b^2+c^2)$ for any $a, b, c \ge 0$. Taking a supremum over n and using the assumptions (41) and (42) yields

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}[\|(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))\|^2] < \infty. \tag{98}$$

Therefore, the weak law of large numbers implies that

$$\widehat{\rho}_n - \overline{\Sigma}_n\beta \xrightarrow{\mathcal{L}_n}_p 0, \tag{99}$$

from which the statement (95) follows by the assumed convergence $\overline{\Sigma}_n \to \overline{\Sigma}$.

□

Lemma 5. *In the setting of Theorem 4, define*

$$\boldsymbol{Y}' \equiv \bar{g}_n(\boldsymbol{Z}) + \boldsymbol{\epsilon}, \quad S_n'^2 \equiv \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2\Sigma_n(\boldsymbol{Z})], \quad \textit{and} \quad \mathcal{L}_n' \equiv \mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Y}', \boldsymbol{Z}). \tag{100}$$

Under the assumptions of Theorem 4(a) or 4(b),

$$\textit{there exist } c_1, c_2 > 0 \textit{ such that } \mathcal{L}_n, \mathcal{L}_n' \in \mathscr{L}(c_1, c_2). \tag{101}$$

Under the assumptions of Theorem 4(a), we have

$$\inf_n \ \lambda_{\min}(S_n^{-1}) > 0, \tag{102}$$

while under the assumptions of Theorem 4(b), we have

$$\lim_{n\to\infty} S_n^2 = \lim_{n\to\infty} S_n'^2 = \overline{\Sigma}^{1/2}(\sigma^2 I_d + \mathcal{E}^2)\overline{\Sigma}^{1/2}. \tag{103}$$

Proof. First, we show that under the assumptions of Theorem 4(a) or 4(b), we have $\mathcal{L}_n \in \mathscr{L}(c_1, c_2)$ for some $c_1, c_2 > 0$. It suffices to show that

$$\sup_n \|S_n^{-1}\| < \infty \tag{104}$$

and

$$\sup_n \ \mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] < \infty. \tag{105}$$

To show the statement (104), first note that

$$\begin{aligned}
S_n^2 &= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n + \boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^2 \Sigma_n(\boldsymbol{Z})] \\
&\quad + 2\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n (\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad + \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^2 \Sigma_n(\boldsymbol{Z})] + \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})],
\end{aligned} \tag{106}$$

where in the last step we used the fact that

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n (\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_n(\boldsymbol{Z}))|\boldsymbol{Z}]^T \beta_n (\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] = 0.
\end{aligned}$$

Furthermore,

$$\begin{aligned}
S_n'^2 &= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{Y}' - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{\epsilon} + \bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{\epsilon}^2 \Sigma_n(\boldsymbol{Z})] + \mathbb{E}_{\mathcal{L}_n}[2\boldsymbol{\epsilon}(\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))\Sigma_n(\boldsymbol{Z})] \\
&\quad + \mathbb{E}_{\mathcal{L}_n}[(\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^2 \Sigma_n(\boldsymbol{Z})] \\
&= \sigma^2 \overline{\Sigma}_n + \overline{\Sigma}_n^{1/2} \mathcal{E}_n^2 \overline{\Sigma}_n^{1/2} = \overline{\Sigma}_n^{1/2} (\sigma^2 I_d + \mathcal{E}_n^2) \overline{\Sigma}_n^{1/2}.
\end{aligned} \tag{107}$$

It follows that $S_n^2 \succcurlyeq \sigma^2 \overline{\Sigma}_n$, which together with assumption (40) implies that

$$\sup_n \|S_n^{-1}\| \leq \sup_n \|\sigma^{-1} \overline{\Sigma}_n^{-1/2}\| = \sigma^{-1} \left(\sup_n \|\overline{\Sigma}_n^{-1}\|\right)^{1/2} < \infty. \tag{108}$$

This verifies statement (104). To prove statement (105), we write

$$\begin{aligned}
&\mathbb{E}_{\mathcal{L}_n}\left[(\boldsymbol{Y} - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad = \mathbb{E}_{\mathcal{L}_n}\left[((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n + \boldsymbol{\epsilon} + \bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad \leq C\mathbb{E}_{\mathcal{L}_n}\left[(((\boldsymbol{X} - \mu_n(\boldsymbol{Z}))^T \beta_n)^4 + \boldsymbol{\epsilon}^4 + (\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4) \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4 | \boldsymbol{Z}]\right] \\
&\quad \leq C\|\beta_n\|^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^8] + 3C\sigma^4 \mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4] \\
&\qquad + C\mathbb{E}_{\mathcal{L}_n}\left[(\bar{g}_n(\boldsymbol{Z}) - \widehat{g}_n(\boldsymbol{Z}))^4 \|\boldsymbol{X} - \mu_n(\boldsymbol{Z})\|^4\right].
\end{aligned}$$

Here, C a constant such that $(a+b+c)^4 \leq C(a^4+b^4+c^4)$ for all $a,b,c \geq 0$. Taking a supremum over n and using the moment assumptions (41) and (42) along with the boundedness of the sequence β_n yields the statement (105).

Therefore, we have verified that $\mathcal{L}_n \in \mathscr{L}(c_1,c_2)$ for some c_1,c_2 under the assumptions of Theorem 4(a) or 4(b). The fact that $\mathcal{L}_n' \in \mathscr{L}(c_1,c_2)$ under these assumptions follows by a similar argument (omitted for the sake of brevity), which finishes the proof of statement (101).

Next, we turn to proving the claim (102). Using calculations (106) and (107), we write

$$S_n^2 = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X}-\mu_n(\boldsymbol{Z}))^T\beta)^2\Sigma_n(\boldsymbol{Z})] + \overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}. \tag{109}$$

Note that

$$\begin{aligned}
&\sup_n \ \|\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X}-\mu_n(\boldsymbol{Z}))^T\beta)^2\Sigma_n(\boldsymbol{Z})]\| \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^2|\boldsymbol{Z}]] \\
&\quad = \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^2|\boldsymbol{Z}]^2] \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^4|\boldsymbol{Z}]] \\
&\quad \leq \sup_n \ \|\beta\|^2\mathbb{E}_{\mathcal{L}_n}[\|\boldsymbol{X}-\mu_n(\boldsymbol{Z})\|^4] < \infty,
\end{aligned} \tag{110}$$

the last step using the eighth moment bound (41). Furthermore,

$$\sup_n \ \|\overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}\| < \infty \tag{111}$$

because $\overline{\Sigma}_n^{1/2}(\sigma^2 I_d + \mathcal{E}_n^2)\overline{\Sigma}_n^{1/2}$ is a convergent sequence by assumption. Hence, $\sup_n \|S_n^2\| < \infty$ and therefore

$$\inf_n \lambda_{\min}(S_n^{-1}) = \inf_n \|S_n\|^{-1} = \inf_n \|S_n^2\|^{-1/2} = \left(\sup_n \|S_n^2\|\right)^{-1/2} > 0.$$

This completes the proof of claim (102).

Finally, we turn to proving claim (103). The claimed convergence of $S_n'^2$ follows immediately from the derivation (107) and the assumption (43). To show that S_n^2 has the same limit, note that the derivation (106) implies that

$$S_n^2 - S_n'^2 = \mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X}-\mu_n(\boldsymbol{Z}))^T\beta_n)^2\Sigma_n(\boldsymbol{Z})] = \frac{1}{n}\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X}-\mu_n(\boldsymbol{Z}))^T h_n)^2\Sigma_n(\boldsymbol{Z})].$$

The boundedness of the quantity $\mathbb{E}_{\mathcal{L}_n}[((\boldsymbol{X}-\mu_n(\boldsymbol{Z}))^T h_n)^2\Sigma_n(\boldsymbol{Z})]$ follows by an argument analagous to that in equation (110), which shows that

$$S_n^2 - S_n'^2 \to 0.$$

This completes the proof of statement (103), so we are done. □

D Proofs for Section 5

Proof of Theorem 5. Let us denote

$$[X, \widetilde{X}]_? \equiv (\{X_j, \widetilde{X}_j\}, X_{-j}, \widetilde{X}_{-j}),$$

where $\{X_j, \widetilde{X}_j\}$ represents the *unordered* pair. In other words, $[X, \widetilde{X}]_?$ specifies $[X, \widetilde{X}]$ up to a swap, hence the "?" notation:

$$[X, \widetilde{X}]_? = [x, \widetilde{x}]_? \quad \Longleftrightarrow \quad [X, \widetilde{X}] \in \{[x, \widetilde{x}], [x, \widetilde{x}]_{\mathrm{swap}(j)}\}.$$

With this notation, we claim that

$$T_j^{\mathrm{opt}} \in \arg\max_{T_j} \mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \;\middle|\; [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right] \quad (112)$$

for every $([x, \widetilde{x}], y)$ in the set

$$\mathcal{A} \equiv \left\{([x, \widetilde{x}], y) : T_j^{\mathrm{opt}}([x, \widetilde{x}], y) \neq T_j^{\mathrm{opt}}([x, \widetilde{x}]_{\mathrm{swap}(j)}, y)\right\}. \quad (113)$$

The conclusion (55) will follow because for any T_j,

$$\begin{aligned}
&\mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)] \\
&= \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y), X_j \neq \widetilde{X}_j] \\
&= \mathbb{P}[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y), ([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{E}\left[\mathbb{P}\left[T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| [X, \widetilde{X}]_?, Y\right] \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&\leq \mathbb{E}\left[\mathbb{P}\left[T_j^{\mathrm{opt}}([X, \widetilde{X}], Y) > T_j^{\mathrm{opt}}([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y) \middle| [X, \widetilde{X}]_?, Y\right] \middle| ([X, \widetilde{X}], Y) \in \mathcal{A}\right] \mathbb{P}[([X, \widetilde{X}], Y) \in \mathcal{A}] \\
&= \mathbb{P}\left[T_j^{\mathrm{opt}}([X, \widetilde{X}], Y) > T_j^{\mathrm{opt}}([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)\right].
\end{aligned}$$

The first step holds because $T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\mathrm{swap}(j)}, Y)$ implies that $X_j \neq \widetilde{X}_j$, the second by the assumption (54), the third and fourth by probability manipulations, the fifth by the claimed conditional optimality (112), and the sixth by the same logic as the first four steps.

To prove equation (112), fix $([x, \widetilde{x}], y) \in \mathcal{A}$. Consider the simple hypothesis testing problem

$$H_0 : (X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \quad \text{versus} \quad H_1 : (X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j), \quad (114)$$

where $(X_j, \widetilde{X}_j)$ are endowed with their law conditional on

$$([X, \widetilde{X}]_?, Y) = ([x, \widetilde{x}]_?, y).$$

We seek the most powerful test of level $\alpha = 1/2$. Note that under the null distribution, the knockoff exchangeability property makes both events equally likely: $\mathbb{P}_0[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j)] = \mathbb{P}_0[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j)] = 1/2$. Therefore, given any statistic T_j, the level $1/2$ test of the simple hypothesis (114) rejects when $T_j([X, \widetilde{X}], Y) > T_j([X, \widetilde{X}]_{\text{swap}(j)}, Y)$. The knockoff statistic T^{opt} defined in equation (112) thus coincides with the most powerful test for the hypothesis (114), which by Neyman-Pearson is given by

$$\begin{aligned}
&T_j^{\text{opt}}([x, \widetilde{x}], y) \\
&\quad = \frac{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right]}{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?, Y = y\right]} \\
&\quad = \frac{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (x_j, \widetilde{x}_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?\right] \mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]\right]}{\mathbb{P}\left[(X_j, \widetilde{X}_j) = (\widetilde{x}_j, x_j) \middle| [X, \widetilde{X}]_? = [x, \widetilde{x}]_?\right] \mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]_{\text{swap}(j)}\right]} \\
&\quad = \frac{\mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]\right]}{\mathbb{P}\left[Y = y \middle| [X, \widetilde{X}] = [x, \widetilde{x}]_{\text{swap}(j)}\right]} = \frac{\mathbb{P}\left[Y = y \,|X_j = x_j, X_{-j} = x_{-j}\right]}{\mathbb{P}\left[Y = y \,|X_j = \widetilde{x}_j, X_{-j} = x_{-j}\right]}.
\end{aligned}$$

The first step is given by Neyman-Pearson, the second by an application of Bayes rule, the third by the conditional exchangeability of knockoffs (50), and the last by the conditional independence of knockoffs (51). Finally, it is easy to verify that

$$\begin{aligned}
&T_j^{\text{opt}}([X, \widetilde{X}], Y) > T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y) \quad \Longleftrightarrow \\
&\mathbb{P}[Y = y | X_j = x_j, X_{-j} = x_{-j}] > \mathbb{P}[Y = y | X_j = \widetilde{x}_j, X_{-j} = x_{-j}],
\end{aligned}$$

from which we conclude that the likelihood given in equation (53) is optimal for the problem (56). This completes the proof. □

Proof of Proposition 1. Suppose $\boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}}$ has a density with respect to the Lebesgue measure. Since

$$\begin{aligned}
&\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] \\
&\quad = \mathbb{E}[\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j} \mid X_{\bullet,-j}, Y, \widetilde{X}]],
\end{aligned}$$

it suffices to show that

$$\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y) \mid X_{\bullet,-j}, Y, \widetilde{X}] = 0$$

for all $X_{\bullet,-j}, Y, \widetilde{X}_j$. Since $\mathcal{L}(\boldsymbol{X}_j | \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}})$ has a density with respect to the Lebesgue measure, so do $\mathcal{L}(\boldsymbol{X}_j | \boldsymbol{Y}, \boldsymbol{X}_{-j}, \widetilde{\boldsymbol{X}})$ and $\mathcal{L}(X_j | Y, X_{\bullet,-j}, \widetilde{X})$. Therefore, it suffices to show that the set

$$S(c; x_{\bullet,-j}, y) \equiv \{x_{\bullet,j} : \mathbb{P}(Y = y | X_{\bullet,j} = x_{\bullet,j}, X_{\bullet,-j} = x_{\bullet,-j}) = c\} \subseteq \mathbb{R}^n$$

has Lebesgue measure zero for all $c, x_{\bullet,-j}, y$. To see this, note that if $x_{\bullet,j} \in S(c; x_{\bullet,-j}, y)$, then

$$\begin{aligned} c &= \mathbb{P}(Y = y | X_{\bullet,j} = x_{\bullet,j}, X_{\bullet,-j} = x_{\bullet,-j}) \\ &= \prod_{i=1}^{n} \exp(\eta_i y_i - \psi(\eta_i)) g_0(y_i) \\ &= \exp\left(\sum_{i=1}^{n} (x_{ij}\beta_j + f_{-j}(x_{i,-j}))y_i - \psi(x_{ij}\beta_j + f_{-j}(x_{i,-j})) + \log g_0(y_i) \right). \end{aligned}$$

It follows that

$$\begin{aligned} &S(c; x_{\bullet,-j}, y) \\ &= \left\{ x_{\bullet,j} : \sum_{i=1}^{n} [x_{ij}\beta_j y_i - \psi(x_{ij}\beta_j + f_{-j}(x_{i,-j}))] = \log c - \sum_{i=1}^{n} [f_{-j}(x_{i,-j})y_i + \log g_0(y_i)] \right\}. \end{aligned} \tag{115}$$

Since ψ is strictly convex and $\beta_j \neq 0$, the left hand side is a strictly concave function of $x_{\bullet,j}$, while the right hand side is a constant (with respect to $x_{\bullet,j}\beta_j$). Thus, $S(c; x_{\bullet,-j}, y)$ is the level set of a strictly concave function, and hence has measure zero. Indeed, the level set of a strictly convex function is the boundary of the corresponding super-level set (which must be convex), and the boundary of any convex set has measure zero [**Lang1986**]. Thus, the conclusion (54) thus follows.

Now, assume that g_η has a density with respect to Lebesgue measure. Since

$$\begin{aligned} &\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}] \\ &\quad = \mathbb{E}[\mathbb{P}[T_j^{\text{opt}}([X, \widetilde{X}], Y) = T_j^{\text{opt}}([X, \widetilde{X}]_{\text{swap}(j)}, Y), X_{\bullet,j} \neq \widetilde{X}_{\bullet,j} \mid X, \widetilde{X}]], \end{aligned}$$

it suffices to show that

$$\mathbb{P}[P(Y|X_{\bullet,j}, X_{\bullet,-j}) = P(Y|\widetilde{X}_{\bullet,j}, X_{\bullet,-j}) \mid X, \widetilde{X}] = 0 \tag{116}$$

for all $X_{\bullet,j} \neq \widetilde{X}_{\bullet,j}$. From expression (115), we see that $P(Y|X_{\bullet,j}, X_{\bullet,-j}) = P(Y|\widetilde{X}_{\bullet,j}, X_{\bullet,-j})$ if and only if

$$\underbrace{\beta_j(X_{\bullet,j} - \widetilde{X}_{\bullet,j})^T}_{\text{slope}} Y - \underbrace{\psi(\beta_j X_{i,j} + f_{-j}(X_{i,-j})) + \psi(\beta_j \widetilde{X}_{i,j} + f_{-j}(X_{i,-j}))}_{\text{intercept}} = 0.$$

Since $\beta_j \neq 0$ by assumption, the slope $\beta_j(X_{\bullet,j} - \widetilde{X}_{\bullet,j}) \neq 0$ and therefore, the set $\{Y : P(Y|X_{\bullet,j}, X_{\bullet,-j}) = P(Y|\widetilde{X}_{\bullet,j}, X_{\bullet,-j})\}$ is a hyperplane (and hence has Lebesgue measure zero). Together with the fact that Y has a density with respect to Lebesgue measure, this implies the relation (116), so the conclusion (54) follows. □