Reconciling model-X and doubly robust approaches to conditional independence testing

Ziang Niu*[1], Abhinav Chakraborty*[1], Oliver Dukes[2], Eugene Katsevich[1] ✉

November 23, 2022

Abstract

Model-X approaches to testing conditional independence between a predictor and an outcome variable given a vector of covariates usually assume exact knowledge of the conditional distribution of the predictor given the covariates. Nevertheless, model-X methodologies are often deployed with this conditional distribution learned in sample. We investigate the consequences of this choice through the lens of the distilled conditional randomization test (dCRT). We find that Type-I error control is still possible, but only if the mean of the outcome variable given the covariates is estimated well enough. This demonstrates that the dCRT is doubly robust, and motivates a comparison to the generalized covariance measure (GCM) test, another doubly robust conditional independence test. We prove that these two tests are asymptotically equivalent, and show that the GCM test is in fact optimal against (generalized) partially linear alternatives by leveraging semiparametric efficiency theory. In an extensive simulation study, we compare the dCRT to the GCM test. We find that the GCM test and the dCRT are quite similar in terms of both Type-I error and power, and that post lasso based test statistics (as compared to lasso based statistics) can dramatically improve Type-I error control for both methods.

1 Introduction

Given a predictor $\boldsymbol{X} \in \mathbb{R}$, response $\boldsymbol{Y} \in \mathbb{R}$, and high-dimensional covariate vector $\boldsymbol{Z} \in \mathbb{R}^p$ drawn from a joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}_n$ (potentially varying with n to accommodate growing p), consider testing the hypothesis of conditional independence (CI)

$$H_{0n} : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{X} \mid \boldsymbol{Z} \tag{1}$$

at level $\alpha \in (0, 1)$ using n data points

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

*These authors contributed equally to this work.

[1]Department of Statistics and Data Science, University of Pennsylvania

[2]Department of Applied Mathematics, Computer Science and Statistics, Ghent University

✉ekatsevi@wharton.upenn.edu

In a high-dimensional regression setting, H_{0n} is a model-agnostic way of formulating the null hypothesis that the predictor $\boldsymbol{X}$ is unimportant in the regression of $\boldsymbol{Y}$ on $(\boldsymbol{X}, \boldsymbol{Z})$ (Candès et al., 2018). In a causal inference setting with treatment $\boldsymbol{X}$, outcome $\boldsymbol{Y}$, observed confounders $\boldsymbol{Z}$, and no unobserved confounders, H_{0n} is the null hypothesis of no causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ (Pearl, 2009).

Formally, a sequence of tests $\phi_n : (X, Y, Z) \mapsto [0, 1]$ of H_{0n} has asymptotic Type-I error control over this null if

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0} \mathbb{E}_{\mathcal{L}_n}[\phi_n(X, Y, Z)] \leq \alpha, \tag{3}$$

where

$$\mathscr{L}_n^0 \equiv \{\mathcal{L}_n : \mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Y} \mid \boldsymbol{Z}) = \mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z}) \times \mathcal{L}_n(\boldsymbol{Y} \mid \boldsymbol{Z})\} \tag{4}$$

is the set of laws satisfying conditional independence. As Shah and Peters (2020) showed, however, in most cases such tests ϕ_n are powerless. The null $\mathscr{L}_n^0$ is therefore too large, and we must content ourselves with controlling Type-I error on only a subset $\mathscr{L}_n^0 \cap \mathscr{R}_n$ where $\mathcal{L}_n$ satisfies additional regularity conditions defined by $\mathscr{R}_n$. We may hope to find nontrivial tests ϕ_n satisfying the following slightly weaker guarantee:

$$\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n(X, Y, Z)] \leq \alpha. \tag{5}$$

One choice of $\mathscr{R}_n$ is $\{\mathcal{L}_n : \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = \mathcal{L}_n^*(\boldsymbol{X}|\boldsymbol{Z})\}$, where $\mathcal{L}_n^*(\boldsymbol{X}|\boldsymbol{Z})$ is a fixed, known distribution. Restriction to $\mathscr{R}_n$, i.e. assuming exact knowledge of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$, is called the *model-X (MX) assumption* (Candès et al., 2018). These authors propose the *conditional randomization test* (CRT) in this setting, which controls Type-I error not just asymptotically (5) but in finite samples as well. The CRT is based on constructing a null distribution for any test statistic by resampling X conditionally on Z using the known conditional law $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. While the CRT is in general computationally costly, using a test statistic of the form

$$T_n^{\mathrm{dCRT}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))$$

gives a fast and powerful test called the *distilled CRT* (dCRT; Liu et al., 2022). Here,

$$\mu_{n,x}(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] \quad \text{and} \quad \mu_{n,y}(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]; \tag{6}$$

the former is known under the MX assumption and the latter is learned in sample. Variants of the dCRT have now been deployed in genetics (Bates et al., 2020) and genomics (Barry et al., 2021) applications.

One of the primary challenges in the practical application of MX methods like the CRT is to obtain the required conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. Outside the context of randomized controlled experiments (Aufiero and Janson, 2022; Ham, Imai, and Janson, 2022), the MX assumption is an approximation (Barber, Candès, and Samworth, 2020;

Huang and Janson, 2020; Li and Liu, 2022). In genome-wide association studies, a realistic parametric distribution can be postulated for this conditional law (Sesia, Sabatti, and Candès, 2019), but the parameters of this distribution must still be learned from data. In cases where a large external unsupervised dataset of $(\boldsymbol{X}, \boldsymbol{Z})$ pairs is available, it has been shown that the distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ can be learned accurately enough for the CRT to control Type-I error (Berrett et al., 2020). In practice, however, the conditional law is usually fit in sample on the same data that is used for testing (Candès et al., 2018; Sesia, Sabatti, and Candès, 2019; Sesia et al., 2020; Bates et al., 2020; Liu et al., 2022; Li et al., 2021; Sesia et al., 2021; Barry et al., 2021). In the case of the dCRT, this results in a procedure we call $\widehat{\text{dCRT}}$ (Algorithm 1) based on the following test statistic:

$$T_n^{\widehat{\text{dCRT}}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)). \tag{7}$$

Algorithm 1: The $\widehat{\text{dCRT}}$.

Input: Data (X, Y, Z), number of randomizations M.

Learn $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ based on (X, Z) and $\widehat{\mu}_{n,y}(\boldsymbol{Z})$ based on (Y, Z);

Compute $T_n^{\widehat{\text{dCRT}}}(X, Y, Z)$;

for $m = 1, 2, \ldots, M$ **do**

  Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^{n} \widehat{\mathcal{L}}_n(X_i|Z_i)$ and compute

$$T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)); \tag{8}$$

end

Output: $\widehat{\text{dCRT}}$ p-value $\frac{1}{M+1}(1 + \sum_{i=1}^{M} \mathbb{1}\{T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z) \geq T_n^{\widehat{\text{dCRT}}}(X, Y, Z)\})$.

The hat on $\widehat{\text{dCRT}}$ is to emphasize that $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is approximated rather than known a priori; this is the procedure that is typically applied in practice instead of the dCRT (which required known $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$). The resampled test statistics $T_n^{\widehat{\text{dCRT}}}(\widetilde{X}^{(m)}, X, Y, Z)$ (8) have four arguments instead of three in order to emphasize that the conditional mean $\widehat{\mu}_{n,x}(\cdot)$ is not refit upon resampling.

The goal of this paper is to study the properties of model-X methods that learn $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ in sample, through the lens of the $\widehat{\text{dCRT}}$. In addition to the quality of the approximation to $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$, we find that the choice of test statistic, and in particular the quality of the estimate $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ is essential: Type-I error is lost when this estimate is poor (Section 3 here and Li and Liu, 2022) but can be guaranteed via a double robustness phenomenon if the estimate is good enough (Section 4). This observation contrasts with the conventional wisdom that the choice of test statistic used with model-X methods impacts power (Katsevich and Ramdas, 2022) but not Type-I error control (Candès et al., 2018). Furthermore, it brings model-X methods more in line with double regression / double machine learning methods (Shah and Peters, 2020; Chernozhukov et al., 2018). This motivates us to compare the $\widehat{\text{dCRT}}$ with the GCM test (Shah and Peters, 2020); we find that the two are in fact asymptotically equivalent (Section 4). Furthermore, we show using semiparametric efficiency theory that the GCM test has optimal power among conditional independence tests against local (generalized) partially linear alternatives (Section 5). Comparing the GCM test to the $\widehat{\text{dCRT}}$ in an extensive simulation study (Section 6), we find that the GCM test usually performs quite similarly to the $\widehat{\text{dCRT}}$ in terms of robustness and power. Since the former test is resampling-free, this suggests it may be preferable in practice to the latter. This work leads to a number of conclusions at the intersection of conditional independence testing, model-X inference, and doubly robust inference (Section 7). In proving our theoretical results, we collate a number of conditional analogs of classical convergence theorems (some but not all novel; Appendix B) and state a slightly sharpened theorem on semiparametric testing optimality (Appendix E), which may be of independent interest.

2 $\widehat{\text{dCRT}}$ resampling distribution converges to normal

To make it easier to analyze the asymptotic properties of the $\widehat{\text{dCRT}}$, in this section we prove that it is asymptotically equivalent to the resampling-free $\widehat{\text{MX(2)}}$ F-test, a variant of the MX(2) F-test (Katsevich and Ramdas, 2022) where the first two moments of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ are estimated in sample. This equivalence was already shown by these authors in the case when $\mu_{n,x}$ is known and $\widehat{\mu}_{n,y}$ is fit out of sample (see their Theorem 2). They conjectured that the equivalence continues to hold when $\widehat{\mu}_{n,y}$ is fit in sample. Here, we prove this conjecture, not just when $\widehat{\mu}_{n,y}$ is fit in sample, but also when the first two moments of $\mu_{n,x}$ are unknown and also fit in sample.

Note that the variance of the resampling distribution of $T_n^{\widehat{\text{dCRT}}}$ is

$$(\widehat{S}_n^{\widehat{\text{dCRT}}})^2 \equiv \text{Var}_{\widehat{\mathcal{L}}_n}[T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z] = \frac{1}{n}\sum_{i=1}^{n} \text{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2. \quad (9)$$

It will be convenient to reformulate $\widehat{\text{dCRT}}$ as

$$\begin{aligned}\phi_n^{\widehat{\text{dCRT}}}(X, Y, Z) &\equiv \mathbb{1}(T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > \mathbb{Q}_{1-\alpha}[T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z]) \\ &= \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > \mathbb{Q}_{1-\alpha}\left[\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z\right]\right) \\ &\equiv \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\text{dCRT}}}} T_n^{\widehat{\text{dCRT}}}(X, Y, Z) > C_n^{\widehat{\text{dCRT}}}(X, Y, Z)\right).\end{aligned}$$

Note that this test is obtained from that in Algorithm 1 by sending $M \to \infty$; we focus our theoretical analysis here and throughout on this infinite-resamples limit of the $\widehat{\text{dCRT}}$.

Here, the α conditional quantile $\mathbb{Q}_\alpha[W \mid \mathcal{F}]$ of a random variable W given a σ-algebra $\mathcal{F}$ is defined via

$$\mathbb{Q}_\alpha[W \mid \mathcal{F}] \equiv \inf\{t : \mathbb{P}[W \leq t \mid \mathcal{F}] \geq \alpha\}. \tag{10}$$

One would expect, based on the central limit theorem, that the conditional distribution of the ratio $T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z)/\widehat{S}_n^{\widehat{\mathrm{dCRT}}}$ tends to $N(0,1)$. This statement is complicated by the conditioning event, which requires us to be careful to define conditional convergence in distribution:

Definition 1. For each n, let W_n be a random variable and let $\mathcal{F}_n$ be a σ-algebra. Then, we say W_n converges in distribution to a random variable W conditionally on $\mathcal{F}_n$ if

$$\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{P}[W \leq t] \text{ for each } t \in \mathbb{R} \text{ at which } t \mapsto \mathbb{P}[W \leq t] \text{ is continuous.} \tag{11}$$

We denote this relation via $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$.

Based on an extension of the Lyapunov central limit theorem to conditional convergence in distribution (Theorem 8), we get the following result:

Theorem 1. [Proof] *Suppose the sequences of true and learned laws $\mathcal{L}_n$ and $\widehat{\mathcal{L}}_n$ satisfy the following two nondegeneracy properties:*

$$\mathbb{P}_{\mathcal{L}_n}[(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2 \geq c] \to 1 \textit{ for some } c > 0; \tag{NDG1}$$

$$0 < \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i], (Y_i - \widehat{\mu}_{n,y}(Z_i))^2, (Y_i - \mu_{n,y}(Z_i))^2 < \infty \textit{ almost surely.} \tag{NDG2}$$

If the conditional Lyapunov condition

$$\frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}\left[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z\right] \xrightarrow{p} 0 \tag{Lyap-1}$$

is satisfied, then

$$\frac{1}{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}} T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z \xrightarrow{d,p} N(0,1) \tag{12}$$

and therefore

$$C_n^{\widehat{\mathrm{dCRT}}}(X, Y, Z) \equiv \mathbb{Q}_{1-\alpha}\left[\frac{1}{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}} T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z) \mid X, Y, Z\right] \xrightarrow{p} z_{1-\alpha}. \tag{13}$$

This suggests that the $\widehat{\mathrm{dCRT}}$ is asymptotically equivalent to the $\widehat{\mathrm{MX(2)}}$ F-test, defined

$$\phi_n^{\widehat{\mathrm{MX(2)}}}(X, Y, Z) \equiv \mathbb{1}\left(\frac{1}{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}} T_n^{\widehat{\mathrm{dCRT}}}(X, Y, Z) > z_{1-\alpha}\right). \tag{14}$$

Indeed, we have the following corollary.

Corollary 1. [Proof] *Consider a sequence of laws $\mathcal{L}_n$ satisfying the assumptions* (NDG1), (NDG2), *and* (Lyap-1) *of Theorem 1, and assume that the test statistic does not accumulate near $z_{1-\alpha}$, i.e.*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\ \mathbb{P}_{\mathcal{L}_n}[|T_n^{\widehat{\text{dCRT}}}(X,Y,Z) - z_{1-\alpha}| \leq \delta] = 0. \tag{15}$$

Then, the $\widehat{\text{dCRT}}$ is asymptotically equivalent to the $\widehat{\text{MX(2)}}$ F-test:

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) = \phi_n^{\widehat{\text{MX(2)}}}(X,Y,Z)] = 1. \tag{16}$$

This result extends Katsevich and Ramdas (2022, Theorem 2) by allowing $\widehat{\mu}_{n,x}$ and $\widehat{\mu}_{n,y}$ to be fit in sample, rather than assuming $\mu_{n,x}$ is known and $\widehat{\mu}_{n,y}$ is fit out of sample. It is a first indication that the dCRT approximates a test based on asymptotic normality.

3 $\widehat{\text{dCRT}}$ is not robust for general $\widehat{\mu}_{n,y}$

One of the hallmarks of MX inference is that it requires "no restriction on the dimensionality of the data or the conditional distribution of $[\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})]$" (Candès et al., 2018). For the CRT, this means that Type-I error is controlled in finite samples, regardless of the test statistic used or the distribution of the response variable. If $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is described by a parametric model with k parameters and we have $N \gg n \cdot k$ unlabeled samples to learn this model, then at least asymptotic Type-I error control is still possible without assumptions on $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$ (Berrett et al., 2020). By contrast, in this section we show that when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is approximated in sample, we cannot expect Type-I error control without assumptions on the response variable.

Let us consider a simple null model $\mathcal{L}_n$ with

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, I_p), \quad \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1), \quad \text{and} \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1). \tag{17}$$

Now, consider a stylized shrinkage estimator $\widehat{\beta}_n \equiv (1 - \frac{c}{\sqrt{n}})\beta$ and a $\widehat{\text{dCRT}}$ based on $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\widehat{\beta}_n, 1)$ and $\widehat{\mu}_{n,y}(\boldsymbol{Z}) \equiv 0$. In this case, the normality of $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ leads to normality of the resampling distribution holding not just asymptotically (12) but in finite samples as well. Therefore, the $\widehat{\text{dCRT}}$ is equal to the $\widehat{\text{MX(2)}}$ F-test:

$$\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) = \mathbb{1}\left(\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - Z_i^T\widehat{\beta}_n)Y_i > z_{1-\alpha}\right). \tag{18}$$

On the other hand, it is easy to derive that

$$\frac{1}{\sqrt{\frac{1}{n}\sum_{i=1}^n Y_i^2}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - Z_i^T\widehat{\beta}_n)Y_i \xrightarrow{d} N\left(\frac{c\|\beta\|^2}{\sqrt{\|\beta\|^2+1}}, 1\right). \tag{19}$$

Therefore, the limiting Type-I error of the $\widehat{\text{dCRT}}$ in this case is

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z)] = 1 - \Phi\left(z_{1-\alpha} - \frac{c\|\beta\|^2}{\sqrt{\|\beta\|^2+1}}\right), \tag{20}$$

which can be made arbitrarily close to one as $c \to \infty$.

Numerical simulations confirm this phenomenon. We constructed a numerical simulation based on the null model (17) with $n = 1600$, $p = 400$, and β having only $s = 5$ nonzero entries (see Section 6.2 below for more on our data-generating model). In this setting, we applied the $\widehat{\text{dCRT}}$ using the lasso and intercept-only models to estimate $\mu_{n,x}$ and $\mu_{n,y}$, respectively. As we increased the magnitude of the coefficient vector β, this test exhibited significant loss of Type-I error control (Figure 1). By contrast, using the lasso instead of the intercept-only model to estimate $\mu_{n,y}$ reduced the Type-I error to nearly the nominal level.

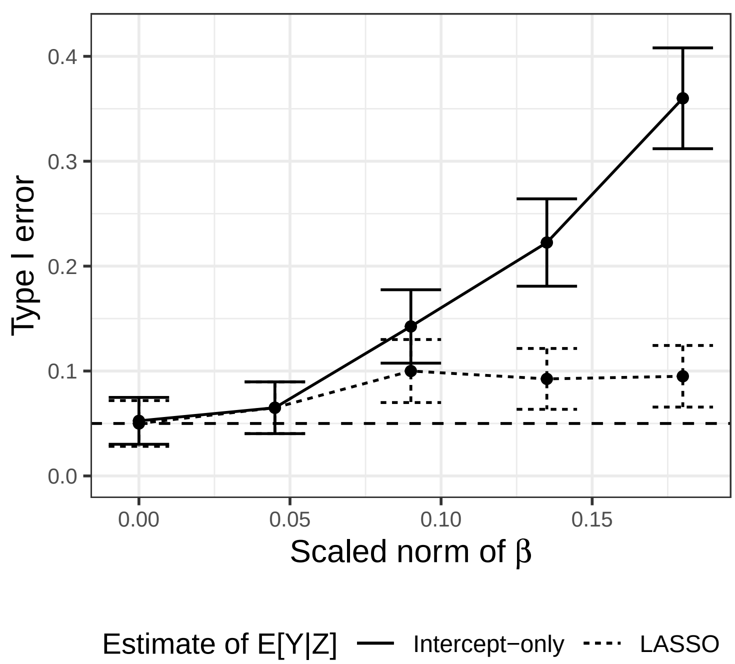

Figure 1: The Type-I error of two instances of the $\widehat{\text{dCRT}}$ under the data-generating model (17), depending on which method is used to estimate $\mu_{n,y}$, when the lasso is used to estimate $\mu_{n,x}$. Improved estimation of $\mu_{n,y}$ leads to markedly reduced Type-I error.

So even when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is estimated at a parametric rate, the $\widehat{\text{dCRT}}$ can have inflated Type-I error rate for certain test statistics. A similar observation was made by Li and Liu (2022) (see the discussion after Theorem 3). Chernozhukov et al. (2018) also give a counterexample in a similar spirit but different context (see the subsection of the introduction titled “regularization bias”). Inspecting the derivation above, we see the issue stems from the fact that $\mathbb{E}_{\mathcal{L}_n}[(\mu_{n,x}(\boldsymbol{Z}) - \widehat{\mu}_{n,x}(\boldsymbol{Z}))(\mu_{n,y}(\boldsymbol{Z}) - \widehat{\mu}_{n,y}(\boldsymbol{Z}))] = O(1/\sqrt{n})$, a rate insufficient for Type-I error control. If we had at least consistency of $\widehat{\mu}_{n,y}(\boldsymbol{Z})$, then this rate would improve to $o(1/\sqrt{n})$ and Type-I error control would be restored. This counterexample suggests that, if $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is learned in sample, then assumptions must be placed not only on $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ but also on $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$ for Type-I error control. This motivates us to investigate the double robustness of the $\widehat{\text{dCRT}}$ and compare it to the GCM test.

4 $\widehat{\text{dCRT}}$ is doubly robust and equivalent to GCM test

Of course, in practice $\widehat{\mu}_{n,y}$ is not fit as naively as in the counterexample from Section 3. The conditional mean $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ is usually approximated via a machine learning algorithm, as improved approximation of this quantity improves the power of the dCRT (Katsevich and Ramdas, 2022). In the context where $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ must be approximated, we claim that more accurate estimation of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ can improve not just the power but also the Type-I error control of the $\widehat{\text{dCRT}}$. This is due to a *double robustness* property of the $\widehat{\text{dCRT}}$, which we state in this section. This property is a consequence of the fact that, under the null, the $\widehat{\text{dCRT}}$ is asymptotically equivalent to another doubly-robust conditional independence test, the GCM test (Shah and Peters, 2020). This equivalence also implies that the $\widehat{\text{dCRT}}$ and GCM test have the same asymptotic power against contiguous alternatives.

4.1 The GCM test

The GCM test is defined as

$$\phi_n^{\text{GCM}}(X,Y,Z) \equiv \mathbb{1}(T_n^{\text{GCM}}(X,Y,Z) > z_{1-\alpha}), \tag{21}$$

where

$$T_n^{\text{GCM}}(X,Y,Z) \equiv \frac{1}{\widehat{S}_n^{\text{GCM}}}\frac{1}{\sqrt{n}}\sum_{i=1}^n (X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)) = \frac{1}{\widehat{S}_n^{\text{GCM}}} T_n^{\widehat{\text{dCRT}}}(X,Y,Z) \tag{22}$$

and $(\widehat{S}_n^{\text{GCM}})^2$ is the empirical variance of the product-of-residual summands:

$$(\widehat{S}_n^{\text{GCM}})^2 \equiv \widehat{\text{Var}}\{(X_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))\}. \tag{23}$$

It controls Type-I error if the following in-sample mean-squared error quantities are small (Shah and Peters, 2020):

$$E_{n,x} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2\right)^{1/2}; \; E'_{n,x} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \text{Var}_{\mathcal{L}_n}[Y_i|Z_i]\right)^{1/2};$$

$$E_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2\right)^{1/2}; \; E'_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2 \text{Var}_{\mathcal{L}_n}[X_i|Z_i]\right)^{1/2}.$$

In particular, Shah and Peters (2020) require that

$$E_{n,x}E_{n,y} = o_{\mathcal{L}_n}(n^{-1/2}),\ E'_{n,x} = o_{\mathcal{L}_n}(1),\ E'_{n,y} = o_{\mathcal{L}_n}(1), \tag{SP1}$$

and, for some constants $c_1, c_2, \delta > 0$,

$$\begin{aligned} &\inf_n\ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))^2(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))^2] > c_1 \\ &\sup_n\ \mathbb{E}_{\mathcal{L}_n}[|(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))|^{2+\delta}] < c_2. \end{aligned} \tag{SP2}$$

The requirement (SP1) that $E_{n,x}E_{n,y} = o_{\mathcal{L}_n}(n^{-1/2})$ is the doubly-robust property of the GCM test: poorer estimates for $\mu_{n,x}$ can be made up for by better estimate of $\mu_{n,y}$.

In the following sections, it will be convenient to augment the assumption (SP1) as follows:

$$E_{n,x}E_{n,y} = o_{\mathcal{L}_n}(n^{-1/2}),\ E'_{n,x} = o_{\mathcal{L}_n}(1),\ E'_{n,y} = o_{\mathcal{L}_n}(1),\ \widehat{E}'_{n,y} = o_{\mathcal{L}_n}(1), \tag{SP1'}$$

where

$$\widehat{E}'_{n,y} \equiv \left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i))^2 \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i]\right)^{1/2}. \tag{24}$$

4.2 Equivalence between GCM test and $\widehat{\text{dCRT}}$

When comparing the GCM test (21) to the $\widehat{\text{MX(2)}}$ F-test (14), which is asymptotically equivalent to the $\widehat{\text{dCRT}}$ (Corollary 1), the only difference is the normalization term. Under the null hypothesis, this difference vanishes asymptotically as long as the estimated variance $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X}|\boldsymbol{Z}]$ is consistent in the following sense:

$$\frac{1}{n}\sum_{i=1}^{n}(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i] - \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i])\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0. \tag{25}$$

Theorem 2. [Proof] *Suppose $\mathcal{L}_n \in \mathscr{L}_n^0$ is a sequence of laws satisfying the assumptions* (SP1') *and* (SP2)*, the nondegeneracy condition* (NDG2)*, the variance consistency property* (25) *and the Lyapunov condition*

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}\left[|Y_i - \mu_{n,y}(Z_i)|^{2+\delta} \mid Z_i\right]\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z] \xrightarrow{p} 0. \tag{Lyap-2}$$

Then, the $\widehat{\text{dCRT}}$ and GCM variance estimates are asymptotically equivalent:

$$\frac{(\widehat{S}_n^{\widehat{\text{dCRT}}})^2}{(\widehat{S}_n^{\text{GCM}})^2} \xrightarrow{p} 1, \tag{26}$$

as are the $\widehat{\text{dCRT}}$ and GCM tests themselves:

$$\lim_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) = \phi_n^{\text{GCM}}(X,Y,Z)] = 1. \tag{27}$$

The variance consistency property (25) is relatively easy to achieve, given the other assumptions of Theorem 2. The following proposition states two sufficient conditions for this property.

Proposition 1. [Proof] *If the assumptions of Theorem 2 other than variance consistency* (25) *hold, then the latter property holds in the following two cases:*

1. $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i] \equiv (X_i - \widehat{\mu}_{n,x}(Z_i))^2$*;*
2. $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[\boldsymbol{X}|\boldsymbol{Z}] \equiv f(\widehat{\mu}_{n,x}(\boldsymbol{Z}))$*, if*
   - $\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}] = f(\mu_{n,x}(\boldsymbol{Z}))$ *for f being Lipschitz on domain* $\cup_{n=1}^{\infty}\mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$ *and* $\mathrm{supp}(\widehat{\mu}_{n,x}(\boldsymbol{Z})) \subseteq \mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$ *almost surely for every n;*
   - $\sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty$.

The first variance estimate given in the proposition can always be applied; the second applies to cases when the mean-variance relationship for $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is known and Lipschitz on the convex hull of the support of $\boldsymbol{X}$, denoted $\mathrm{Conv}(\mathcal{L}_n(\boldsymbol{X}))$. This is the case, for example, if $\boldsymbol{X}$ is binary and we define $f(t) \equiv t(1-t)$.

One consequence of Theorem 2 is that the $\widehat{\mathrm{dCRT}}$ and GCM test are also asymptotically equivalent against local alternatives, so in particular have the same power.

Corollary 2. [Proof] *If $\mathcal{L}_n'$ is a sequence of alternative distributions that is contiguous to a sequence $\mathcal{L}_n \in \mathscr{L}_n^0$ satisfying the assumptions of Theorem 2, then the $\widehat{\mathrm{dCRT}}$ and GCM tests are asymptotically equivalent against $\mathcal{L}_n'$:*

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n'}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) = \phi_n^{\mathrm{GCM}}(X,Y,Z)] = 1 \tag{28}$$

and therefore have the same asymptotic power:

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_n'}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}_n'}[\phi_n^{\mathrm{GCM}}(X,Y,Z)] = 0. \tag{29}$$

By constructing a null distribution via resampling, the CRT allows for arbitrarily complicated test statistics whose asymptotic distributions are not known. For the $\widehat{\mathrm{dCRT}}$, however, the resampling-based null distribution simply recapitulates the asymptotic normal distribution used by the GCM test (Theorems 1 and 2). Therefore, at least in large samples, the extra computational burden of resampling is unnecessary as the equivalent GCM can be applied instead.

4.3 Double robustness of $\widehat{\mathrm{dCRT}}$

Another consequence of Theorem 2 is that the $\widehat{\mathrm{dCRT}}$ is doubly robust under the variance consistency condition (25), since it is equivalent under the null hypothesis to the doubly robust GCM test.

Corollary 3. [Proof] *Let $\mathscr{R}_n$ be a sequence of regularity conditions such that for any sequence $\mathcal{L}_n \in \mathscr{R}_n$, we have the nondegeneracy condition* (NDG2)*, the Lyapunov condition* (Lyap-2)*, the conditions* (SP1') *and* (SP2)*, and consistent variance estimates* (25)*. Then, the $\widehat{\text{dCRT}}$ has asymptotic Type-I error control over $\mathscr{L}_n^0 \cap \mathscr{R}_n$ in the sense of the definition* (5)*.*

Therefore, Type-I error control requires accuracy of only the first two moments of $\widehat{\mathcal{L}}_n$, in parallel to Theorem 2 of Katsevich and Ramdas (2022). The condition on the second moment of $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ is needed because the variance of the resampling distribution must not be smaller (asymptotically) than the true variance of the test statistic. This condition does not require much more than accurate estimation of the first moments (Proposition 1). It can be dropped altogether if we build normalization directly into the $\widehat{\text{dCRT}}$ test statistic. We explore this possibility in Appendix A.

Our conclusion that $\widehat{\text{dCRT}}$ is doubly robust appears at odds with the statement that "the model-X CRT...does not pursue such double robustness through learning and adjusting for both $X|Z$ and $Y|Z$..." (Li and Liu, 2022). The CRT, and the dCRT in particular, *do* usually learn both $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$; the former is learned when approximating the model for X and the latter when computing the test statistic. If the quality of these estimates is sufficiently good, then dCRT will control Type-I error (Corollary 3). We assume what Li and Liu (2022) meant is that if the estimate for $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$ is poor, then we cannot expect the dCRT to control Type-I error when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is learned in sample. We agree with this statement, as discussed in Section 3.

5 GCM test is optimal against certain alternatives

We have shown that, in large samples, the dCRT has the same power against local alternatives as the resampling-free GCM test. Of course, other instances of the much more general CRT paradigm have better power than the GCM test against certain alternatives. We show in this section, however, that this is not the case for generalized partially linear models (GPLMs), a broad class of alternatives. In fact, the GCM test is asymptotically most powerful against GPLM alternatives. We leverage classical semiparametric efficiency theory (Choi, Hall, and Schick, 1996; Van Der Vaart, 1998; Kosorok, 2008) to prove this result. We state our optimality result in Section 5.1, give an example of its application in Section 5.2, and then compare it to existing semiparametric optimality results in Section 5.3.

5.1 Optimality result

To facilitate the link with semiparametric theory, in this section of the paper we operate in a fixed-dimensional setting. Accordingly, we drop the subscript n from $\mathscr{L}_n^0$ and $\mathscr{R}_n$. For each value of n, we have $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \in \mathbb{R}^{1+1+p}$ for fixed p. We will seek power against semiparametric GPLM alternatives of the form

$$\mathcal{L}_\theta(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \equiv \mathcal{L}_{\beta,\eta}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \equiv \mathcal{L}_{x,z}(\boldsymbol{X}, \boldsymbol{Z}) \times f_{\boldsymbol{\eta}}(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}), \quad \boldsymbol{\eta} = \boldsymbol{X}\beta + g(\boldsymbol{Z}). \tag{30}$$

Here, f_η is a one-parameter exponential family with natural parameter $\eta \in \mathbb{R}$ and log-partition function ψ, $\beta \in \mathbb{R}$ and

$$g \in \mathcal{H}_g \subseteq L^2(\mathcal{L}_{x,z}(\boldsymbol{Z})), \tag{31}$$

where $\mathcal{H}_g$ is a linear subspace of the L^2 space of functions on $\mathbb{R}^p$ with the measure $\mathcal{L}_{x,z}(\boldsymbol{Z})$. The alternatives (30) are those where $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}$ follows an exponential family distribution with natural parameter linear in $\boldsymbol{X}$ and potentially nonlinear in $\boldsymbol{Z}$. Note that GPLMs include linear and generalized linear models as special cases, and therefore cover a broad range of alternative distributions.

We focus on power against local alternatives $\mathcal{L}_{\theta_n(h)}$ near $\theta_0 \equiv (0, g_0)$, defined by

$$\theta_n(h) \equiv \theta_n(h_\beta, h_g) \equiv (h_\beta/\sqrt{n}, g_0 + h_g/\sqrt{n}), \quad \text{for} \quad h \equiv (h_\beta, h_g) \in (0, \infty) \times \mathcal{H}_g. \tag{32}$$

We leave the dependence of $\theta_n(h)$ on g_0 implicit. Next, we define asymptotic optimality against such local alternatives following Choi, Hall, and Schick (1996):

Definition 2. For $h \in (0, \infty) \times \mathcal{H}_g$, we say a test ϕ_n^* is the locally asymptotically most powerful level α test of

$$H_0 : \mathcal{L} \in \mathscr{R} \subseteq \mathscr{L}^0 \quad \text{versus} \quad H_{1n} : \mathcal{L} = \mathcal{L}_{\theta_n(h)} \tag{33}$$

if ϕ_n^* has asymptotic Type-I error control over $\mathscr{R}$ at level α and for any other test ϕ_n satisfying the same property we have

$$\limsup_{n\to\infty} \ \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n(X, Y, Z)] \le \liminf_{n\to\infty} \ \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n^*(X, Y, Z)]. \tag{34}$$

If this is true for every $h \in (0, \infty) \times \mathcal{H}_g$, such a test is locally asymptotically uniformly most powerful at g_0, or LAUMP(g_0). A test is LAUMP($\mathcal{S}$) against $\mathcal{L}_{\theta_n(h)}$ for $h \in (0, \infty) \times \mathcal{H}_g$ if it is LAUMP(g_0) for each $g_0 \in \mathcal{S} \subseteq \mathcal{H}_g$.

Finally, define

$$s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]]. \tag{35}$$

We are now ready to state our main optimality result.

Theorem 3. [Proof] *Consider the conditional independence testing problem* (33)*, with a collection of null distributions $\mathscr{R} \subseteq \mathscr{L}^0$ satisfying some regularity conditions, a linear subspace $\mathcal{H}_g \subseteq L^2(\mathcal{L}_{x,z}(\boldsymbol{Z}))$ specifying possible values for the nonparametric component g in the GPLM alternative model* (30)*, and some subset $\mathcal{S} \subseteq \mathcal{H}_g$. If the following four assumptions hold:*

$$\textit{assumptions } (\text{SP1}) \textit{ and } (\text{SP2}) \textit{ hold for all } \mathcal{L} \in \mathscr{R}, \tag{36}$$

$$\ddot{\psi} = K > 0 \textit{ and } \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty \textit{ OR } \mathrm{supp}(\boldsymbol{X}, \boldsymbol{Z}) \textit{ is compact and } \mathcal{H}_g \subseteq C(\mathbb{R}^p), \tag{37}$$

$$\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}| \cdot] \in \mathcal{H}_g, \tag{38}$$

$$\forall \ g_0 \in \mathcal{S}, h_g \in \mathcal{H}_g, \ \mathcal{L}_{\theta_n(0,h_g)} \in \mathscr{R} \textit{ for large enough } n, \tag{39}$$

then ϕ_n^{GCM} *is LAUMP($\mathcal{S}$) against* $\mathcal{L}_{\theta_n(h)}$ *for* $h \in (0,\infty) \times \mathcal{H}_g$*, with*

$$\lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n^{\mathrm{GCM}}(X,Y,Z)] = 1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)). \tag{40}$$

Let us discuss each of the four assumptions of Theorem 3:

- The assumption (36) is a set of regularity conditions on the null distributions $\mathscr{R}$. It is the same set of assumptions made by Shah and Peters (2020) to ensure Type-I error control of the GCM test over $\mathscr{R}$, including the assumption that the conditional means $\mu_{n,x}$ and $\mu_{n,y}$ are fit accurately enough (SP1) and fairly mild moment assumptions (SP2).

- The assumption (37) is a set of regularity conditions on the alternative distribution (30). These conditions are required for the semiparametric optimality theory to apply. These assumptions allow for GPLMs based on the normal distribution (assuming $\boldsymbol{X}$ has second moment) or any other exponential family (assuming $(\boldsymbol{X}, \boldsymbol{Z})$ is compactly supported and the functions g are continuous).

- The assumption (38) states that the conditional expectation $\boldsymbol{Z} \mapsto \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\boldsymbol{Z}]$ must belong to the subspace $\mathcal{H}_g$. It guarantees that the "least favorable" value of the nonparametric component g is in the space $\mathcal{H}_g$, yielding the optimality of the GCM statistic.

- The assumption (39) connects the semiparametric alternative hypothesis to the conditional independence null hypothesis. In some sense it requires $\mathcal{L}_{\theta_0} \equiv \mathcal{L}_{(0,g_0)}$ (derived from the semiparametric alternative distribution (30)) to be an interior point of $\mathscr{R}$ (the conditional independence null) for each $g_0 \in \mathcal{S}$.

We give an example of when these assumptions hold in the next section.

5.2 Example: kernel ridge regression

We illustrate Theorem 3 with a kernel ridge regression example, borrowed from Shah and Peters (2020, Section 4). Suppose the conditional expectations $\mu_x(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mu_y(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}}[\boldsymbol{Y}|\boldsymbol{Z}]$ satisfy $\mu_x, \mu_y \in \mathcal{H}_k$ for some reproducing kernel Hilbert space $(\mathcal{H}_k, \|\cdot\|_{\mathcal{H}_k})$ with reproducing kernel $k : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. In particular, we consider $\mathcal{H}_k \equiv W^{1,2}([0,1]) \subset L^2([0,1])$, i.e. the Sobolev space defined

$$W^{1,2}([0,1]) \equiv \big\{f : [0,1] \to \mathbb{R} \mid f(0) = 0,\ f \text{ is absolutely continuous with } \dot{f} \in L^2([0,1])\big\},$$

equipped with the inner product

$$\langle f, g \rangle_{W^{1,2}([0,1])} \equiv \int_0^1 \dot{f}(z)\dot{g}(z)\mathrm{d}z.$$

$W^{1,2}([0,1])$ is an RKHS with kernel $k(x,y) = \min\{x,y\}$ (Wainwright, 2019, Example 12.16). Consider the kernel ridge estimators

$$\begin{aligned}
\widehat{\mu}_x &\equiv \underset{\mu_x \in W^{1,2}([0,1])}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^n |X_i - \mu_x(Z_i)|^2 + \lambda \|\mu_x\|^2_{W^{1,2}([0,1])} \right\}; \\
\widehat{\mu}_y &\equiv \underset{\mu_y \in W^{1,2}([0,1])}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^n |Y_i - \mu_y(Z_i)|^2 + \lambda \|\mu_y\|^2_{W^{1,2}([0,1])} \right\},
\end{aligned} \tag{41}$$

with λ tuned as described in Shah and Peters (2020, Section 4). Using Shah and Peters (2020, Theorem 11), the following result can be derived as a consequence of Theorem 3.

Corollary 4. [Proof] *Fix $C > 0$, and consider the following regularity class $\mathscr{R} \subseteq \mathscr{L}^0$:*

$$\begin{aligned}
\mathscr{R} \equiv \{&\mathcal{L}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z}) = \mathcal{L}(\boldsymbol{Z}) \times \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) \times \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) : \\
&\mathcal{L}(\boldsymbol{Z}) = \mathrm{Unif}([0,1]),\ \mathcal{L}(\boldsymbol{X}|\boldsymbol{Z}) = N(\mu_x(\boldsymbol{Z}),1),\ \mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) = N(\mu_y(\boldsymbol{Z}),1), \\
&\mu_x, \mu_y \in B_{W^{1,2}}(0,C)\},
\end{aligned} \tag{42}$$

where we define the $W^{1,2}([0,1])$ ball

$$B_{W^{1,2}}(0,C) \equiv \{f \in W^{1,2}([0,1]) : \|f\|_{W^{1,2}([0,1])} < C\}. \tag{43}$$

Now, fix $\mu_{0x}, \mu_{0y} \in B_{W^{1,2}}(0,C)$ and for each $h = (h_\beta, h_g) \in (0,\infty) \times W^{1,2}([0,1])$ consider the set of local alternatives $\mathcal{L}_{\theta_n(h)}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})$ given by

$$\begin{aligned}
&\mathcal{L}_{\theta_n(h)}(\boldsymbol{Z}) \equiv \mathrm{Unif}([0,1]); \\
&\mathcal{L}_{\theta_n(h)}(\boldsymbol{X}|\boldsymbol{Z}) \equiv N(\mu_{0x}(\boldsymbol{Z}),1); \\
&\mathcal{L}_{\theta_n(h)}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{Z}) \equiv N(\boldsymbol{X}h_\beta/\sqrt{n} + \mu_{0y}(\boldsymbol{Z}) + h_g(\boldsymbol{Z})/\sqrt{n}, 1).
\end{aligned} \tag{44}$$

Then, the GCM test based on the kernel ridge estimators (41) is LAUMP($B_{W^{1,2}}(0,C)$).

Hence, the GCM test based on kernel ridge regression does not just control Type-I error (Shah and Peters, 2020, Theorem 11); it is also optimal against local alternatives.

5.3 Discussion of Theorem 3

Theorem 3 states that the GCM test of Shah and Peters (2020) is the optimal test of conditional independence against a broad class of semiparametric GPLM alternatives, including linear and generalized linear models. To our knowledge, it is the first result at the intersection of conditional independence testing and semiparametric optimality, although Shah and Peters (2020) have already noted the connection between the GCM test and nonparametric estimation of the expected conditional covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ given $\boldsymbol{Z}$. Our result complements another line of work on minimax optimality for conditional independence testing (Canonne et al., 2018; Neykov, Balakrishnan, and Wasserman, 2021; Kim et al., 2022). In the related model-X context, few optimality results are available. Two existing works show optimality statements based on likelihood ratio statistics; one in the context of the CRT (Katsevich and Ramdas, 2022) and the other in the context of model-X knockoffs (Spector and Fithian, 2022).

Theorem 3 closely parallels results on estimation in semiparametric regression (Robinson, 1988; Bickel et al., 1993; Donald and Newey, 1994; Härdle, Liang, and Gao, 2000; Robins and Rotnitzky, 2001; Van De Geer et al., 2014; Ning and Liu, 2017; Janková and Van De Geer, 2018; Chernozhukov et al., 2018). It is well-known that the GCM statistic with the true conditional means μ_x and μ_y is the efficient score in the context of GPLMs, and, when normalized by the efficient information, gives an efficient estimator of the target parameter. Existing results on semiparametric optimality for hypothesis testing state that tests based on optimal estimators are themselves optimal (Choi, Hall, and Schick, 1996; Van Der Vaart, 1998; Kosorok, 2008).

Despite the similarity between Theorem 3 and existing semiparametric optimality results, we emphasize that this theorem is a statement about optimality for conditional independence testing rather than for semiparametric testing. The semiparametric model (30) plays the role of the alternative distribution with respect to which power is evaluated, and need not hold under the null hypothesis. To bridge this gap, it suffices to find an open ball within the conditional independence null hypothesis containing the semiparametric null hypothesis (39). This allows us to reduce the conditional independence testing problem to a semiparametric testing problem, and therefore to leverage existing semiparametric optimality results (Appendix E).

Note that Theorem 3 gives the power against local alternatives of the GCM test with μ_x and μ_y estimated in sample. This complements Shah and Peters (2020, Theorem 8), where these authors compute the power of the GCM test against non-local alternatives by resorting to sample splitting, which is not required to show Type-I error control for the GCM test. This sample splitting is necessary under non-local alternatives to avoid Donsker conditions; using either sample splitting or Donsker conditions is also standard practice in the semiparametric literature. By contrast, we avoid sample splitting by exploiting the special structure of the conditional independence null and contiguity arguments to compute limiting power under local alternatives.

While the Type-I error control results in Section 4 are stated in the high-dimensional setting, Theorem 3 is stated only for fixed-dimensional covariate vectors $\boldsymbol{Z}$. Indeed, semiparametric optimality theory is predominantly low-dimensional. A notable exception is the work of Janková and Van De Geer (2018), which provides a semiparametric theory of estimation in high dimensions. Extending this theory to hypothesis testing is nontrivial, and beyond the scope of the current work. Nevertheless, proving optimality statements for conditional independence testing in high dimensions is an interesting direction for future work. We note in passing that high-dimensional results for lasso-based estimators often assume exact sparsity of the coefficient vector, which poses a problem for conditions (39) requiring the regularity class $\mathscr{R}$ to have interior points.

Finally, we note that Theorem 3 gives the optimality of the GCM statistic against alternative models for $\boldsymbol{Y}$ in which $\boldsymbol{X}$ and $\boldsymbol{Z}$ do not interact. For alternatives where the conditional association between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is modified by $\boldsymbol{Z}$, the GCM test will no longer be optimal. Variants of the CRT (Zhong, Kuffner, and Lahiri, 2021; Sesia and Sun, 2022), model-X knockoffs (Li et al., 2021), and the GCM test (Lundborg et al., 2022) are designed to improve power in the presence of effect modification are available, although their optimality properties are not described. Optimal tests developed specifically for detecting interaction effects between $\boldsymbol{X}$ and $\boldsymbol{Z}$ (rather than main effects) may be constructed based on Vansteelandt et al. (2008).

6 Finite-sample performance assessment

The results in the preceding sections are all asymptotic. In this section, we complement these results with a comprehensive simulation-based assessment of Type-I error and power in finite samples. Previous simulation-based assessments of the Type-I error of MX methods have come to differing conclusions: Sesia, Sabatti, and Candès (2019), Romano, Sesia, and Candès (2019), Sesia et al. (2020), and Liu et al. (2022) found broad robustness to misspecification of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ while Li and Liu (2022) found such misspecifications to cause marked Type-I error inflation. We show that differences in the level of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ implied by the simulation design explain these discrepancies, and then use this insight to inform our own simulation design in Section 6.2. Then, we present the results of our numerical simulations in Section 6.3. Numerical simulation results and instructions to reproduce them are available at `https://github.com/Katsevich-Lab/symcrt-manuscript-v1`.

6.1 Revisiting prior simulations of robustness

The question of robustness of MX methods to the misspecification of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ has been investigated starting from the paper in which the model-X framework was originally proposed (Candès et al., 2018). In this paper, the joint distribution $\mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Z})$ was estimated in sample via the graphical lasso, which is similar to estimating the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ via the ordinary lasso. These authors found that

> "Although the graphical Lasso is well suited for this problem since the covariates have a sparse precision matrix, its covariance estimate is still off by nearly 50%, and yet surprisingly the resulting power and FDR are nearly indistinguishable from when the exact covariance is used...the nominal level of 10% FDR is never violated, even for covariance estimates very far from the truth."

Similar conclusions have been drawn from numerical simulations in subsequent papers as well (Sesia, Sabatti, and Candès, 2019; Romano, Sesia, and Candès, 2019; Sesia et al., 2020; Liu et al., 2022), the latter studying the dCRT specifically. On the other hand, the numerical simulations of Li and Liu (2022) show that the dCRT can suffer significant Type-I error inflation when $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ is inaccurately fit. These authors state that "for model-X inference, the dependence of $\boldsymbol{X}$ on $\boldsymbol{Z}$ is not adequately characterized and adjusted [for] due to the shrinkage bias of lasso."

To resolve this apparent contradiction, we consider a common data-generating model used in MX literature:

$$\mathcal{L}_n(\boldsymbol{X}, \boldsymbol{Z}) = N(0, \Sigma), \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}^T\theta + \boldsymbol{Z}^T\beta, \sigma_y^2). \tag{45}$$

Often, $(\boldsymbol{X}, \boldsymbol{Z})$ are assumed to have a spatial structure (motivated by the GWAS application), with $\Sigma = \Sigma(\rho) \in \mathbb{R}^{(1+p)\times(1+p)}$ taken to be the AR(1) covariance matrix with autocorrelation parameter $\rho \in (-1, 1)$. This covariance matrix roughly approximates linkage disequilibrium structure among genotypes, where correlations among variables are local with respect to the spatial structure. Conditional independence under this model (45) reduces to $H_0 : \theta = 0$. Furthermore, the conditional distribution $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ implied by the normal joint distribution is that of a linear model:

$$\text{Under } H_0, \quad \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\gamma, \sigma_x^2), \quad \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, \sigma_y^2). \tag{46}$$

In the context of this model, the conditional independence testing problem is nontrivial to the extent that $\boldsymbol{Z}$ induces marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ even in the absence of conditional association. In a causal inference context, this spurious marginal association would be called a confounding effect of $\boldsymbol{Z}$. This marginal association can be small or large, depending on the correlation structure of $\boldsymbol{Z}$ and the extent to which the supports of β and γ overlap. Properly adjusting for $\boldsymbol{Z}$ is important to the extent that $\boldsymbol{Z}$ induces marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$.

We claim that the simulation studies in much of the original MX literature had relatively low levels of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$, whereas the simulation studies in Li and Liu (2022) were done in a regime with much more marginal association. To illustrate this point, we quantify the level of marginal association in a given problem setup as the Type-I error of the GCM test with intercept-only models for $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ and $\mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{Z})$. This test is essentially a Pearson test of (marginal) independence between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and ignores the variables $\boldsymbol{Z}$ altogether. We compute this Type-I error for the data-generating models used to assess robustness by Candès et al. (2018), Liu et al. (2022), and Li and Liu (2022) (Appendix F.1). The former two papers are framed in the variable selection context, where several explanatory variables $\boldsymbol{W}_j$ are considered, and the hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{\text{-}j}$ is tested for each j. Therefore, $\boldsymbol{X} \equiv \boldsymbol{W}_j$ for each j. On the other hand, Li and Liu (2022) considered a conditional independence testing framework, where $\boldsymbol{X}$ was a single variable of interest.

For the data-generating models used by Candès et al. (2018) and Liu et al. (2022), we evaluate the Type-I error of the marginal GCM test for each hypothesis $H_0 : \boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{W}_j \mid \boldsymbol{W}_{\text{-}j}$, plotting these as a function of j (Figure 2, top row). We superimpose onto these plots a blue horizontal line indicating the Type-I error of the marginal GCM test for the data-generating model used by Li and Liu (2022) (equal to 0.99, suggesting strong marginal association), and a red dashed horizontal line indicating the nominal level of this marginal test (equal to 0.05). The green ticks indicate the locations of the non-null variables. As expected for a setting where variable correlation is local, we see that Type-I error is inflated for null variables near the signal variables. The extent of this inflation depends on the autocorrelation parameter (set at 0.3 by Candès et al., 2018 and 0.5 by Liu et al., 2022) and the locations of the signal variables. Most null variables, however, are not near signal variables, and therefore the marginal GCM test shows no inflation. This is reflected by the histograms of the Type-I error inflations (Figure 2, bottom row). The median Type-I error of the marginal GCM test is near the nominal level of 0.05 in all three of the simulation setups from Candès et al. (2018) and Liu et al. (2022).

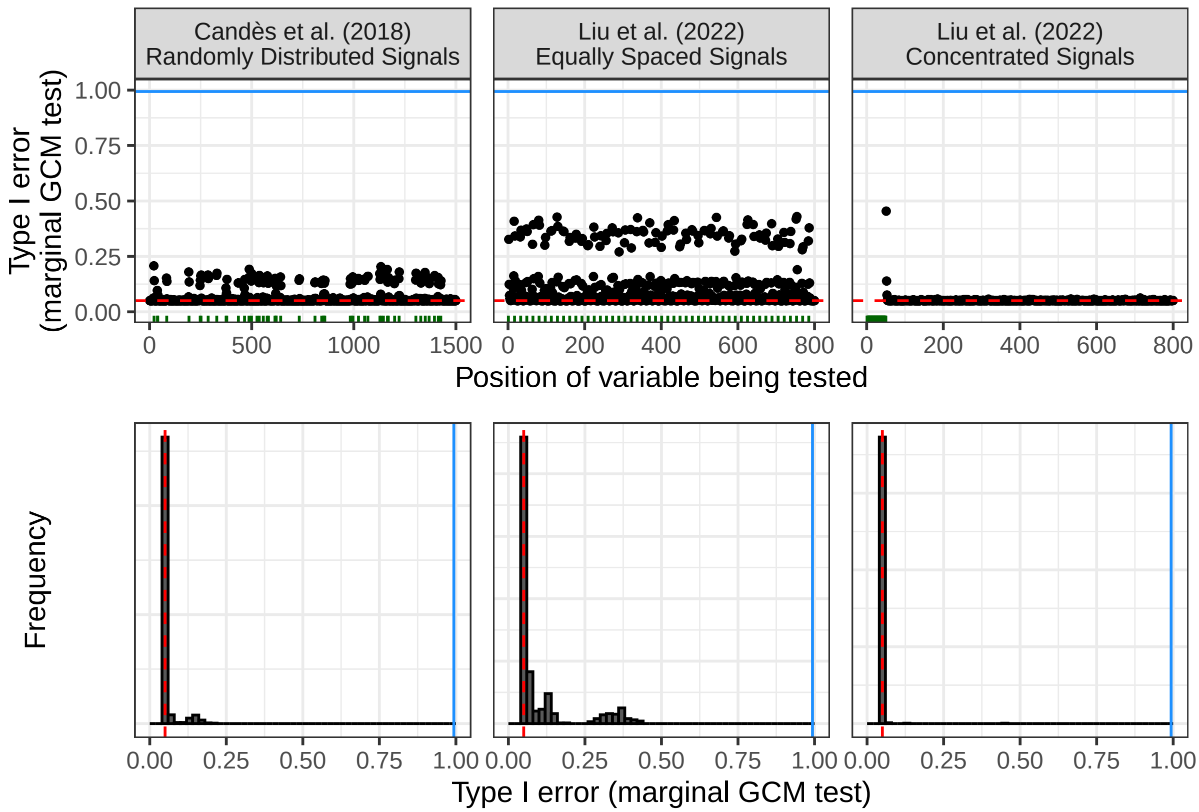

Figure 2: Comparing the marginal associations between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the robustness simulations of Candès et al. (2018), Liu et al. (2022), and Li and Liu (2022) (Appendix F.1). Top: Type-I error of the marginal GCM test as a function of the position of null variables with respect to the non-null variables (represented as green ticks). Bottom: Histograms of the Type-I error across null variables. The solid blue line indicates the Type-I error of the marginal GCM test for the robustness simulation of Li and Liu (2022), and the dashed red line the nominal Type-I error level of the marginal GCM test (0.05).

6.2 Simulation design

Data-generating model. As discussed in the previous section, appropriately setting the marginal correlation between $\boldsymbol{X}$ and $\boldsymbol{Y}$ in a given data-generating model is crucial to properly evaluate the impact of inaccurate estimation of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ on the Type-I error control of a model-X method. Keeping this in mind, we propose the following data-generating model:

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, \Sigma(\rho)),\ \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1),\ \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\boldsymbol{X}^T\theta + \boldsymbol{Z}^T\beta, 1). \quad (47)$$

We set the first s coefficients of β to be equal to ν and the rest to zero. Therefore, the entire data-generating process is parameterized by the six parameters $(n, p, s, \rho, \theta, \nu)$ (Table 1). For both null and alternative simulations, we vary each of the first four across five values each, setting the remaining three to the default value indicated in bold. The fifth parameter θ controls the signal strength and the sixth parameter ν controls the extent of marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$. For the null simulation, we set $\theta \equiv 0$, and for each setting of (n, p, s, ρ), we choose five values of ν equally spaced between 0 (no marginal association) and $\nu_{\max}$ (computed so that the marginal GCM method has Type-I error 0.99). Note that $\nu_{\max}$ depends on the parameters (n, p, s, ρ), so not exactly the same values of ν were used across settings of these four parameters. For the alternative simulation, we kept ν fixed at $\nu_{\max}/2$ while for each setting of (n, p, s, ρ), we choose five values of θ equally spaced between 0 (no signal) and $\theta_{\max}$ (computed so that the GCM method with oracle settings of $\widehat{\mu}_{n,x}$ and $\widehat{\mu}_{n,y}$ has power 0.99). Finally, we complement the linear regression data-generating model (47) with an analogous one based on logistic regression.

n	p	s	ρ	θ (null)	ν (null)	θ (alt)	ν (alt)
100	100	**5**	0	0	0	0	$\nu_{\max}/2$
200	200	10	0.2	0	$\nu_{\max}/4$	$\theta_{\max}/4$	$\nu_{\max}/2$
400	**400**	20	**0.4**	0	$\nu_{\max}/2$	$\theta_{\max}/2$	$\nu_{\max}/2$
800	800	40	0.6	0	$3\nu_{\max}/4$	$3\theta_{\max}/4$	$\nu_{\max}/2$
1600	1600	80	0.8	0	$\nu_{\max}$	$\theta_{\max}$	$\nu_{\max}/2$

Table 1: The values of the sample size n, covariate dimension p, autocorrelation of covariates ρ, sparsity s, signal strength θ, and marginal association strength ν used for the simulation study. Each of the parameters n, p, s, ρ was varied among the values in the first table while keeping the other three at their default values, indicated in bold. For example, $p = 400, s = 5, \rho = 0.4$ were kept fixed while varying $n \in \{100, 200, 400, 800, 1600\}$. The second and third tables denote the values of (θ, ν) used for the null and alternative simulations. Each combination of (n, p, s, ρ) was paired with each of the five values of (θ, ν) displayed for null and alternative simulations.

Methodologies compared. In Section 4, we found that the GCM test and the $\widehat{\text{dCRT}}$ are equivalent when applied with the same estimation methods for $\mu_{n,x}$ and $\mu_{n,y}$. Using this equivalence, we also showed that the $\widehat{\text{dCRT}}$ is robust to errors in $\widehat{\mu}_{n,x}$ if they are compensated for by accurate estimates $\widehat{\mu}_{n,y}$. In our simulation to assess Type-I error, we wish to probe the finite-sample Type-I error control of the GCM and the $\widehat{\text{dCRT}}$. We apply both of these methods with the lasso to estimate $\mu_{n,x}$ and $\mu_{n,y}$, as this is the most common choice in the MX literature.

In addition to the GCM test and the $\widehat{\text{dCRT}}$, we apply the Maxway CRT (Li and Liu, 2022), designed specifically to improve the Type-I error control of the dCRT in the context when $\mu_{n,x}$ must be estimated. The Maxway CRT is inherently a semi-supervised method, assuming the existence of an auxiliary unlabeled dataset containing observations of $\boldsymbol{X}$ and $\boldsymbol{Z}$ but not of $\boldsymbol{Y}$. The methodology (specifically, "Maxway$_{\text{in}}$ example 1") proceeds—roughly—by fitting $\widehat{\mathcal{L}}_n(\boldsymbol{X}|\boldsymbol{Z})$ on the unlabeled data via the post-lasso (i.e. selecting active variables via the lasso and then refitting via ordinary least squares, Belloni and Chernozhukov, 2013), fitting $\widehat{\mu}_{ny}(\boldsymbol{Z})$ on the labeled data via post-lasso, and then applying dCRT on the labeled data based on these two models.

Since the primary focus of this paper is the setting when no auxiliary unlabeled data are available, we implement the Maxway CRT by randomly splitting the data into two equal pieces, using the first as the unlabeled data (in particular, ignoring the response data) and the second as the labeled data. This strategy is consistent with the real data analysis in Li and Liu (2022, Section 6). We also consider a bona-fide semi-supervised setup, in order to compare the GCM test and $\widehat{\text{dCRT}}$ to the Maxway CRT in the setting originally considered by Li and Liu (2022). However, in the semi-supervised setting we use all of the available data on $(\boldsymbol{X}, \boldsymbol{Z})$ (i.e. both unlabeled and labeled data) to fit $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$. By contrast, Li and Liu (2022) used only the unlabeled data to learn $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ in their implementation of the $\widehat{\text{dCRT}}$ for semi-supervised data.

Finally, we noted in Section 4 that the $\widehat{\text{dCRT}}$ already has a built-in doubly robust property. Therefore, we conjectured that the Type-I error inflation observed in the simulations of Li and Liu (2022) is attributable to poor estimation of $\mu_n(\boldsymbol{X}|\boldsymbol{Z})$ and/or $\mu_n(\boldsymbol{Y}|\boldsymbol{Z})$ and that the $\widehat{\text{dCRT}}$ can achieve Type-I error control if used in conjunction with better estimators of these conditional means. Taking inspiration from Li and Liu (2022), we also considered versions of the $\widehat{\text{dCRT}}$ and the GCM test based on the post-lasso in addition to those based on the usual lasso. In summary, we compared five methods: lasso and post-lasso based GCM, lasso and post-lasso based $\widehat{\text{dCRT}}$, and Maxway CRT (Table 2). As a point of reference for the null simulation, we also included the GCM test with intercept-only models for $\mu_{n,x}$ and $\mu_{n,y}$; the Type-I error of this test quantifies the degree of marginal association in the data-generating model. As a point of reference for the alternative simulation, we also included the GCM test with $\mu_{n,x}$ and $\mu_{n,y}$ set to their ground truth values; the power of this test is the maximum power achievable by any test and therefore quantifies the signal strength in the data-generating model.

Evaluation of power in the presence of Type-I error inflation. The methodologies compared control Type-I error to differing extents across the variety of simulation parameters in Table 1. This makes it challenging to compare power across methods, since some control Type-I error while others do not. To address this challenge, we chose to compare the power of the *test statistics* underlying the methods, each under oracle calibration to ensure Type-I error control. Given the composite null, exact oracle calibration

Method name	Estimating $\mu_{n,x}$	Data for $\widehat{\mu}_{n,x}$	Estimating $\mu_{n,y}$	Data for $\widehat{\mu}_{n,y}$
GCM (LASSO)	lasso	all	lasso	all/labeled
$\widehat{\text{dCRT}}$ (LASSO)	lasso	all	lasso	all/labeled
GCM (PLASSO)	post-lasso	all	post-lasso	all/labeled
$\widehat{\text{dCRT}}$ (PLASSO)	post-lasso	all	post-lasso	all/labeled
Maxway CRT	post-lasso	unlabeled	post-lasso	labeled
GCM (marginal)	intercept-only	all	intercept-only	all/labeled
GCM (oracle)	ground truth	–	ground truth	–

Table 2: The five methodologies compared, how they estimate $\mu_{n,x}$ and $\mu_{n,y}$, and what data they use for each in the context of semi-supervised or fully supervised data. Note that in the fully supervised case, data is split in half to form "unlabeled" and labeled sets for Maxway CRT. In this case, the $\widehat{\text{dCRT}}$ and GCM tests still use all of the data for estimating $\mu_{n,x}$ and $\mu_{n,y}$. Two additional tests were used for reference purposes: the GCM test with intercept-only models for $\mu_{n,x}$ and $\mu_{n,y}$ and the GCM test with $\mu_{n,x}$ and $\mu_{n,y}$ set to their ground truth values.

is computationally intractable. Therefore, we instead calibrated each test with respect to the point null given by

$$\mathcal{L}_n(\boldsymbol{Z}) = N(0, \Sigma(\rho)),\ \mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z}) = N(\boldsymbol{Z}^T\beta, 1),\ \mathcal{L}_n(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z}) = N(\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]^T\theta + \boldsymbol{Z}^T\beta, 1).$$

This is the "closest" point in the null to the alternative (47) under consideration; therefore ensuring Type-I error control at this point null should be a decent proxy for ensuring Type-I error control over the whole null. To calibrate two-sided tests with respect to this point null, we generate samples of a test statistic from the null and then define lower and upper critical values as the 2.5% and 97.5% quantiles of this distribution. Using potentially asymmetric lower and upper critical values is necessary, as the null distribution may not be symmetric and centered at zero (Liu et al., 2022).

6.3 Simulation results

We conducted simulations for Gaussian and binary models for the response $\boldsymbol{Y}$, each within the supervised and semi-supervised settings. We present the Type-I error and power for Gaussian responses in the supervised setting in Figures 3 and 4, respectively, while deferring the other cases to Appendix F.3. Note also that for the sake of brevity Figures 3 and 4 only present three out of the five values for the four parameters n, p, s, ρ; the complete results are presented in Appendix F.3.

Next we list the main conclusions regarding Type-I error based on the results in Figures 3 (Gaussian supervised), 8 (Gaussian semi-supervised), 10 (binary supervised), and 12 (binary semi-supervised):

- As one would expect, across all simulation settings, all methods have poorer Type-I error control as sample size n decreases, dimension p increases, sparsity s increases, autocorrelation ρ increases, or marginal association strength ν increases.

- For Gaussian responses, the $\widehat{\text{dCRT}}$ and GCM methods based on the same test statistics have very similar Type-I error control, echoing the asymptotic equivalence of the two methods (Theorem 2). For binary responses, the lasso-based $\widehat{\text{dCRT}}$ has somewhat lower Type-I error than the lasso-based GCM test (Figure 10). The discreteness of binary responses likely slows down the convergence to normality of the GCM statistic, rendering the resampling-based null distribution of the $\widehat{\text{dCRT}}$ a better approximation to the null distribution.

- Across all simulation settings, the $\widehat{\text{dCRT}}$ and GCM methods based on the post-lasso have dramatically better Type-I error control than their lasso-based counterparts. This is because the post-lasso tends to more fully regress the confounders $\boldsymbol{Z}$ out of the response $\boldsymbol{Y}$; see also Section F.2 below.

- Across all simulation settings, Maxway CRT has better Type-I error control than the lasso-based $\widehat{\text{dCRT}}$ (in line with the results of Li and Liu, 2022), but worse Type-I error control than the post-lasso-based $\widehat{\text{dCRT}}$. The latter is likely due to the fact that Maxway CRT uses only half of the available data on $(\boldsymbol{X}, \boldsymbol{Z})$ to fit $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$, and therefore does not adjust for $\boldsymbol{Z}$ as accurately.

Next, we list the main conclusions regarding power based on the results in Figures 4 (Gaussian supervised), 9 (Gaussian semi-supervised), 11 (binary supervised), and 13 (binary semi-supervised):

- Across all simulation settings, GCM-based methods have somewhat higher power than their $\widehat{\text{dCRT}}$-based methods. This may have to do with the stabilizing effect of the GCM normalization, compared to the unnormalized $\widehat{\text{dCRT}}$ statistic. The difference between the two tends to vanish as sample size grows, reflecting the asymptotic equivalence of the two methods (Corollary 2).

- Across all simulation settings, the $\widehat{\text{dCRT}}$ and GCM methods based on the lasso have lower power than their post-lasso-based counterparts. This is because the post-lasso introduces more variance into the estimation of $\mu_{n,y}$; see also Section F.2 below.

- Across Gaussian and binary supervised simulation settings (Figures 7 and 11), Maxway CRT has the lowest power among all methods compared. The reason for this is that Maxway CRT relies on data splitting and therefore has half the effective sample size of the other methods. On the other hand, for semi-supervised settings (Figures 9 and 13), Maxway CRT has power comparable to or better than those of the post-lasso-based methods, but still worse than the lasso-based methods. This is due to the additional variance introduced by the refitting step in the post-lasso.

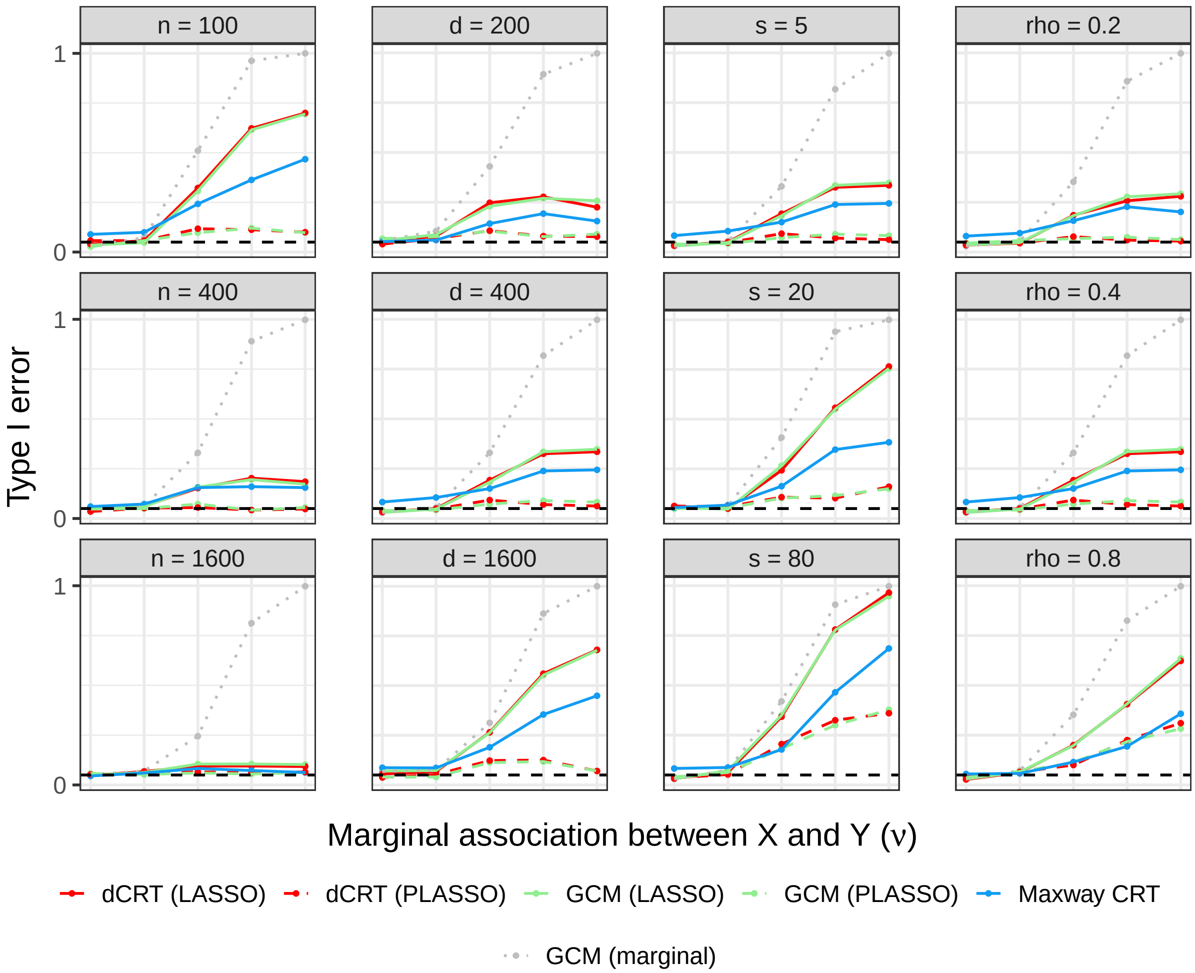

Figure 3: Type I error control for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the marginal association strength ν in each subplot. Each point is the average of 400 Monte Carlo replicates.

In summary, the methods with the best Type-I error control across all simulation settings are the $\widehat{\text{dCRT}}$ and the GCM test based on the post-lasso, although this improved robustness does come with a cost in terms of power when compared to the lasso-based methods. We investigate the associated trade-off in Appendix F.2.

7 Conclusion

Model-X inference with $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ fit in sample can be doubly robust. Model-X inference (Candès et al., 2018) is presented as a mode of inference where the assumptions are transferred entirely from $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z})$ to $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$; no restrictions are made on the former law (or the test statistic used, at least in the context of the CRT), while the latter law

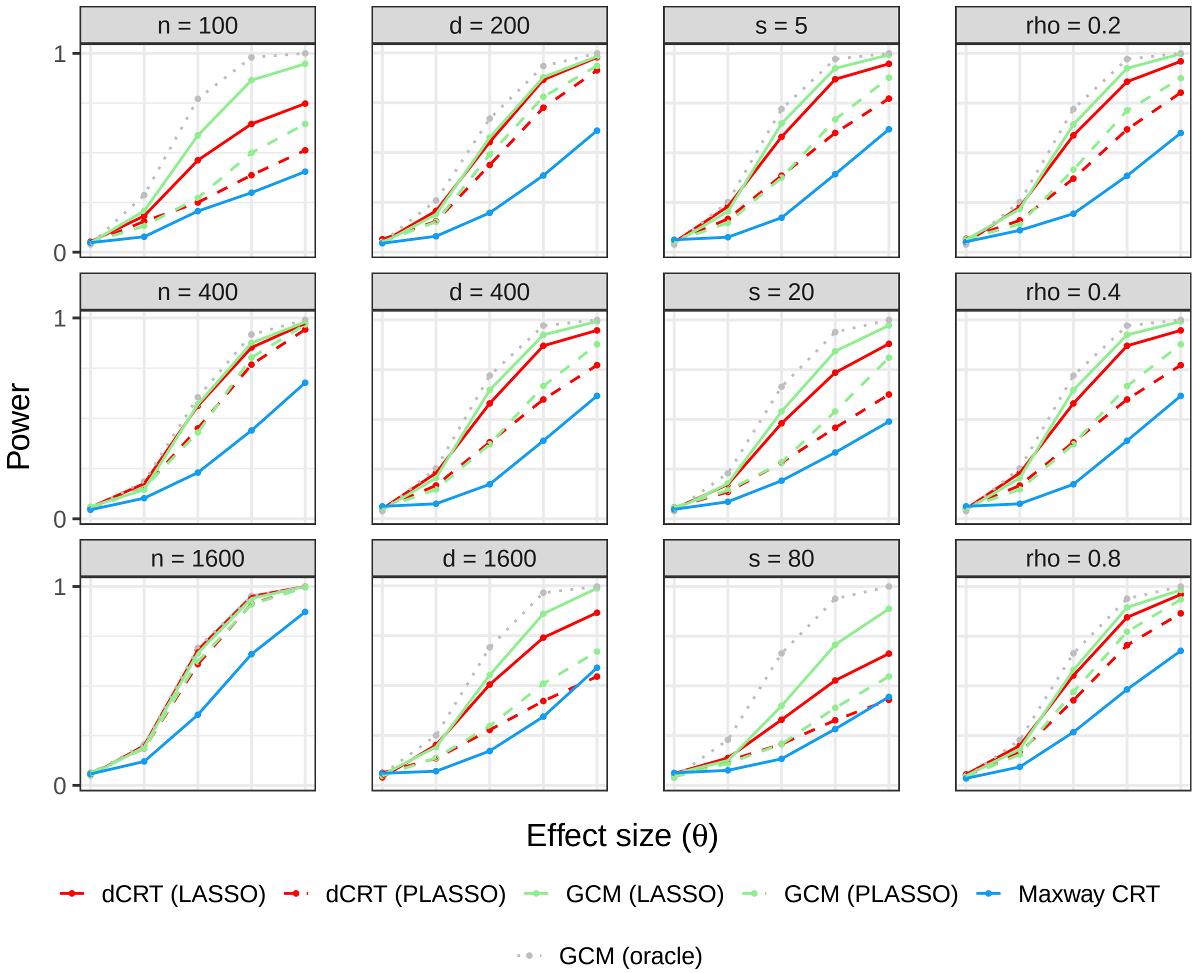

Figure 4: Type I error control for Gaussian supervised setting: we vary only one parameter in each column and there are five values of the signal strength θ in each subplot. Each point is the average of 400 Monte Carlo replicates.

is assumed exactly known. In practice, however, the law $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ is often fit in sample. In the context of the dCRT, we show that Type-I error control cannot be guaranteed without restrictions on $\mathcal{L}(\boldsymbol{Y}|\boldsymbol{Z})$ or the test statistic used (Section 3). On the other hand, test statistics based on decent estimates of $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ can compensate for errors in the estimation of $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ and restore Type-I error control (Corollary 3), a double robustness phenomenon. This result brings model-X inference more in line with double regression inferential methodologies: The conditional mean $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ is estimated in the context of in-sample approximation to the "model for X," and the conditional mean $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ is estimated when computing the model-X test statistic. Relatedly, a double robustness property was noted for conditional model-X knockoffs (Huang and Janson, 2020). A doubly robust version of the dCRT has also been recently proposed (the Maxway CRT; Li and Liu, 2022), although we argue that the original dCRT is itself doubly robust.

The GCM test has similar Type-I error and power as the dCRT. When fitting $\mathcal{L}(\boldsymbol{X}|\boldsymbol{Z})$ in sample, the dCRT is essentially a double regression methodology. This prompts a comparison to the GCM test (Shah and Peters, 2020), another conditional independence test based on double regression. We established that the two tests are asymptotically equivalent under the null (Theorem 2) and under arbitrary local alternatives (Corollary 2). This suggests that the dCRT and the GCM test—when applied with the same estimators for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$—should have similar Type-I error control and power. Our numerical simulations (Section 6) largely confirm this behavior in finite samples. A possible exception to this conclusion is the case when discreteness in the response variable $\boldsymbol{Y}$ slows down the convergence of the $\widehat{\text{dCRT}}$ resampling distribution to normality (Theorem 1). In such cases, we observed that the $\widehat{\text{dCRT}}$ can in fact have better Type-I error control than the GCM based on the same estimators (Figure 10), presumably thanks to a better approximation to the null distribution in finite samples. Nevertheless, the broad similarity between the performance of the GCM test and the dCRT, together with the fact that the former test requires no resampling, suggests that the GCM test may be preferable to the dCRT in many practical problems.

The post lasso yields much better Type-I error control than the lasso. Double robustness results for the GCM test and the dCRT apply only insofar as the estimation methods used in conjunction with these tests are accurate enough (SP1). The default estimation method for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ in many model-X applications is the lasso. As was demonstrated by Li and Liu, 2022, the shrinkage bias of the lasso leads to inadequate adjustment of $\boldsymbol{X}$ and $\boldsymbol{Y}$ for $\boldsymbol{Z}$, which in turn leads to inflated Type-I error. The same authors proposed the Maxway CRT, an extension of the dCRT involving the identification of coordinates of $\boldsymbol{Z}$ impacting $\boldsymbol{X}$ and $\boldsymbol{Y}$ via the lasso followed by least squares refitting. Inspired by this work, we applied the original dCRT with post lasso estimates for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$. We found vastly improved Type-I error control (Figure 6), compared not just to the lasso-based dCRT but also to the Maxway CRT itself. The decreased bias of the post lasso helps adjust for $\boldsymbol{Z}$ more fully, although we found that the extra variance incurred by refitting does come at a cost in power. Nevertheless, our results suggest that applying the post lasso in conjunction with model X methodologies can lead to significant improvements in robustness.

The GCM test is the optimal conditional independence test against alternatives without interactions between $\boldsymbol{X}$ and $\boldsymbol{Z}$. It is widely known in the semiparametric literature that the GCM test is the efficient score test for (generalized) partially linear models. The connection between the GCM test and semiparametric theory was noted briefly by Shah and Peters (2020), though not explored in depth; presumably this is because the GCM test is a conditional independence test rather than a test of a parameter in a semiparametric model. Nevertheless, we find that if the semiparametric *null* hypothesis can be embedded within the conditional independence null hypothesis (39), semiparametric optimality theory can be carried over fairly directly to conditional independence testing to establish optimality against semiparametric *alternative* distributions (Theorem 3). Thanks to this connection, we find that the GCM test has optimal asymptotic power among conditional independence tests against local generalized partially linear model alternatives (30). On the other hand, we leave open the question of optimality against alternatives where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are allowed to interact. We also leave open whether our optimality result can be extended to the high-dimensional regime.

Limitations: The proportional regime and the variable selection problem. Our results about the equivalence between the GCM test and the dCRT, and the double robustness of the latter, require estimates of $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ that are individually consistent and whose rates of convergence are sufficiently fast (SP1). In the case of sparse linear models, we can get such rates if $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$ depend on at most $s = o(\sqrt{n}/\log(p))$ of the coordinates of $\boldsymbol{Z}$. Such assumptions are common in other lines of work on high-dimensional / semiparametric / doubly-robust inference, including the debiased lasso (Van De Geer et al., 2014; Zhang and Zhang, 2014; Javanmard and Montanari, 2014; Ning and Liu, 2017; Janková and Van De Geer, 2018) and doubly-robust causal inference (Belloni, Chernozhukov, and Hansen, 2014; Chernozhukov et al., 2018). On the other hand, consistent estimates are typically not available in the regime when n, p, and s grow proportionally (Bayati and Montanari, 2011), causing a failure in traditional debiased estimates (Celentano and Montanari, 2021). An additional limitation of the current work is that we do not directly consider the variable selection problem. For example, application of the GCM test to each variable is much more computationally costly than applying model-X knockoffs. Therefore, the comparison between model-X and doubly robust methodologies for variable selection purposes requires more thought.

Acknowledgments

EK was partially supported by NSF DMS-2113072. ZN was partially supported by the grant “Statistical Software for Single Cell CRISPR Screens” awarded to EK by Analytics at Wharton. We gratefully acknowledge help from Timothy Barry with our simulation studies and the underlying computational infrastructure, including his `simulatr` R package and Nextflow pipeline. We also acknowledge dedicated support from the staff at the Wharton High Performance Computing Cluster. We also thank Lucas Janson for providing details about the simulation setting in Candès et al. (2018).

A The $\widehat{\text{dCRT}}$ with GCM normalization

As an alternative to the $\widehat{\text{dCRT}}$, we consider the $\widehat{\text{ndCRT}}$. This procedure is based on a normalized statistic that coincides exactly with the GCM statistic:

$$T_n^{\widehat{\text{ndCRT}}}(X, Y, Z) \equiv \frac{1}{\widehat{S}_n^{\text{GCM}}} T_n^{\widehat{\text{dCRT}}}(X, Y, Z) \equiv T_n^{\text{GCM}}(X, Y, Z).$$

The only difference with GCM is that the critical value is given by conditional resampling rather than a normal quantile:

$$\begin{aligned}\phi_n^{\widehat{\text{ndCRT}}}(X,Y,Z) &\equiv \mathbb{1}(T_n^{\widehat{\text{ndCRT}}}(X,Y,Z) > \mathbb{Q}_{1-\alpha}[T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z]) \\ &\equiv \mathbb{1}\left(T_n^{\widehat{\text{ndCRT}}}(X,Y,Z) > C_n^{\widehat{\text{ndCRT}}}(X,Y,Z)\right).\end{aligned} \tag{48}$$

Here,

$$T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \equiv \frac{T_n^{\text{dCRT}}(\widetilde{X},X,Y,Z)}{S_n^{\text{GCM}}(\widetilde{X},X,Y,Z)},$$

where

$$(S_n^{\text{GCM}}(\widetilde{X},X,Y,Z))^2 \equiv \widehat{\text{Var}}\{(\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i))\}$$

and $T_n^{\text{dCRT}}(\widetilde{X},X,Y,Z)$ is as defined in equation (8).

Theorem 4. [Proof] *Let $\mathcal{L}_n$ be a sequence of laws such that the nondegeneracy conditions* (NDG1) *and* (NDG2) *and the conditional Lyapunov condition* (Lyap-1) *hold. Then,*

$$T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z \xrightarrow{d,p} N(0,1) \tag{49}$$

and therefore

$$C_n^{\widehat{\text{ndCRT}}}(X,Y,Z) \equiv \mathbb{Q}_{1-\alpha}\left[T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \mid X,Y,Z\right] \to z_{1-\alpha} \tag{50}$$

and the $\widehat{\text{ndCRT}}$ *and GCM tests are equivalent:*

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{ndCRT}}}(X,Y,Z) = \phi_n^{\text{GCM}}(X,Y,Z)] = 1. \tag{51}$$

Corollary 5. [Proof] *Let $\mathscr{R}_n$ be a sequence of regularity conditions such that for any sequence $\mathcal{L}_n \in \mathscr{R}_n$, we have the the nondegeneracy conditions* (NDG1) *and* (NDG2)*, the conditional Lyapunov condition* (Lyap-1)*, and the assumptions* (SP1) *and* (SP2)*. Then, the* $\widehat{\text{ndCRT}}$ *has asymptotic Type-I error control over $\mathscr{L}_n^0 \cap \mathscr{R}_n$ in the sense of the definition* (5)*.*

Comparing Corollary 5 to Corollary 3, we see that the $\widehat{\text{ndCRT}}$ controls Type-I error under weaker assumptions than the $\widehat{\text{dCRT}}$; in particular the variance of $\mathcal{L}_n(\boldsymbol{X}|\boldsymbol{Z})$ need not be estimated with any degree of accuracy.

B Conditional convergence results

The proofs of our theoretical results rely on the conditional counterparts of several standard convergence theorems. In this section, we state these conditional convergence theorems. We defer their proofs to Appendix G.

First we define a notion of conditional convergence in probability, analogous to our definition of conditional convergence in distribution (Definition 1).

Definition 3. For each n, let W_n be a random variable and let $\mathcal{F}_n$ be a σ-algebra. Then, we say W_n converges in probability to a constant c conditionally on $\mathcal{F}_n$ if W_n converges in distribution to the delta mass at c conditionally on W_n (recall Definition 1). We denote this convergence by $W_n \mid \mathcal{F}_n \xrightarrow{p,p} c$. In symbols,

$$W_n \mid \mathcal{F}_n \xrightarrow{p,p} c \quad \text{if} \quad W_n \mid \mathcal{F}_n \xrightarrow{d,p} \delta_c. \tag{52}$$

Now we are ready to state the conditional convergence results.

B.1 Statements

For the sake of all results below, let $\mathcal{F}_n$ be a sequence of σ-algebras.

Theorem 5 (Conditional Polya's theorem)**.** [Proof] *Let W_n be a sequence of random variables. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W with continuous CDF, then*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| \xrightarrow{p} 0. \tag{53}$$

Theorem 6 (Conditional Slutsky's theorem)**.** [Proof] *Let W_n be a sequence of random variables. Suppose a_n and b_n are sequences of random variables such that $a_n \xrightarrow{p} 1$ and $b_n \xrightarrow{p} 0$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W with continuous CDF, then*

$$a_n W_n + b_n \mid \mathcal{F}_n \xrightarrow{d,p} W. \tag{54}$$

Theorem 7 (Conditional law of large numbers)**.** [Proof] *Let W_{in} be a triangular array of random variables, such that W_{in} are independent conditionally on $\mathcal{F}_n$ for each n. If for some $\delta > 0$ we have*

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] \xrightarrow{p} 0, \tag{55}$$

then

$$\frac{1}{n} \sum_{i=1}^{n} (W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]) \mid \mathcal{F}_n \xrightarrow{p,p} 0. \tag{56}$$

The condition (55) *is satisfied when*

$$\sup_{1\leq i\leq n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] = o_p(n^{\delta}). \tag{57}$$

As a corollary of Theorem 7, if we choose $\mathcal{F}_n = \{\varnothing, \Omega\}$, we are able to obtain the following version of the weak law of large numbers for triangular arrays.

Corollary 6 (Unconditional weak law of large numbers)**.** [Proof] *Let W_{in} be a triangular array of random variables, such that W_{in} are independent for each n. If for some $\delta > 0$ we have*

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta}] \to 0, \tag{58}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(W_{in} - \mathbb{E}[W_{in}]) \xrightarrow{p} 0. \tag{59}$$

The condition (58) *is satisfied when*

$$\sup_{1\le i\le n} \mathbb{E}[|W_{in}|^{1+\delta}] = o(n^{\delta}). \tag{60}$$

Theorem 8 (Conditional central limit theorem)**.** [Proof] *Let W_{in} be a triangular array of random variables, such that for each n, W_{in} are independent conditionally on $\mathcal{F}_n$. Define*

$$S_n^2 \equiv \sum_{i=1}^{n} \operatorname{Var}[W_{in} \mid \mathcal{F}_n], \tag{61}$$

and assume $0 < \operatorname{Var}[W_{in}|\mathcal{F}_n] < \infty$ almost surely for all $i = 1, \dots, n$ and for all $n \in \mathbb{N}$. If for some $\delta > 0$ we have

$$\frac{1}{S_n^{2+\delta}}\sum_{i=1}^{n} \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^{2+\delta} \mid \mathcal{F}_n] \xrightarrow{p} 0, \tag{62}$$

then

$$\frac{1}{S_n}\sum_{i=1}^{n}(W_{in} - \mathbb{E}[W_{in} \mid \mathcal{F}_n]) \mid \mathcal{F}_n \xrightarrow{p} N(0,1). \tag{63}$$

Lemma 1 (Conditional convergence implies quantile convergence)**.** [Proof] *Let W_n be a sequence of random variables and $\alpha \in (0,1)$. If $W_n \mid \mathcal{F}_n \xrightarrow{d,p} W$ for some random variable W whose CDF is continuous and strictly increasing at $\mathbb{Q}_\alpha[W]$, then*

$$\mathbb{Q}_\alpha[W_n \mid \mathcal{F}_n] \xrightarrow{p} \mathbb{Q}_\alpha[W]. \tag{64}$$

B.2 Discussion

The above definitions and results on conditional convergence are not particularly surprising, and related results are present in the existing literature. Nevertheless, we have not found any of the above results stated in the literature in exactly this form. Here we discuss the relationships of our definitions and results with existing ones.

Notions of conditional convergence in probability and in distribution have been explicitly defined by Nowak and Zięba (2005). However, these notions require a single conditioning σ-algebra as well as almost sure convergences of conditional probabilities, whereas in Definitions 1 and 3 we allow the conditioning σ-algebra to change with n and for the conditional probabilities to converge in probability. Our Definition 1 can be viewed as formalizing the notion of conditional convergence in distribution implicitly used by Wang and Janson (2022). Related notions of conditional convergence in distribution allowing for changing conditioning σ-algebra are present implicitly in the works of Dedecker and Merlevede (2002) and Bulinski (2017), though these are based on the convergence of conditional characteristic functions as opposed to conditional cumulative distribution functions.

Turning to the convergence results themselves, we were not able to find conditional Polya's or Slutsky's theorems (Theorems 5 and 6) in the literature. Versions of the conditional law of large numbers are given by Majerek, Nowak, and Zieba, 2005 and Prakasa Rao, 2009, but these involve a single conditioning σ-algebra and do not allow for triangular arrays, unlike Theorem 7. Remarkably, we could not find even the unconditional triangular array law of large numbers (Corollary 6) in the literature; existing results either assume a second-moment condition or use truncation (Durrett, 2010, Theorems 2.2.4 and 2.2.6, respectively) instead of a $1+\delta$ moment condition or are not applicable to triangular arrays (Shah and Peters, 2020, Lemma 19). As for laws of large numbers, Grzenda and Zieba (2008), Prakasa Rao (2009), and Yuan, Wei, and Lei (2014) give non-triangular array versions of the conditional central limit theorem that require a single conditioning σ-algebra, unlike Theorem 8. Versions of the conditional central limit theorem appropriate for varying conditioning σ-algebras and triangular arrays are given by Dedecker and Merlevede (2002) and Bulinski (2017), those these involve different notions of conditional convergence in distribution. Finally, we note that our result that conditional convergence in distribution implies in-probability quantile convergence (Lemma 1) is a generalization of Wang and Janson (2022, Lemma 3) to general conditioning σ-algebras.

C Proofs for Section 2

C.1 Proofs of main results

Proofs of Theorem 1 and Corollary 1. We prove instead the stronger Theorem 9 and Corollary 7 below. $\square$

Theorem 9. *Let $\mathcal{L}_n$ be a sequence of laws and $\widehat{\mathcal{L}}_n$ be a sequence of estimates. Suppose either of the following two sets of assumptions is satisfied:*

1. *The nondegeneracy conditions* (NDG1) *and* (NDG2) *and the conditional Lyapunov condition* (Lyap-1) *holds.*

2. *The assumptions of Theorem 2 hold.*

Then, the normalized $T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z)$ converges conditionally to a normal distribution (12) *and therefore the $\widehat{\mathrm{dCRT}}$ critical value $C_n^{\widehat{\mathrm{dCRT}}}(X, Y, Z)$ converges to $z_{1-\alpha}$* (13).

Proof. It suffices to prove the conditional convergence in distribution (12), as the convergence of the critical value (13) follows from Lemma 1 because the normal distribution has continuous and strictly increasing CDF. We prove the conditional convergence (12) for each of the two sets of assumptions.

Assumption 1. We proceed by applying the conditional CLT (Theorem 8) with

$$W_{in} \equiv \frac{1}{\sqrt{n}}(\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)) \tag{65}$$

and $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. To verify the assumptions of the conditional CLT, note first that W_{in} are independent conditionally on $\mathcal{F}_n$ by construction and satisfy $0 < \mathrm{Var}[W_{in} \mid \mathcal{F}_n] < \infty$ by the nondegeneracy assumption (NDG2). Next, recalling definition (9), we have

$$s_n^2 \equiv \sum_{i=1}^n \mathrm{Var}_{\widehat{\mathcal{L}}_n}[W_{in} \mid X, Y, Z] \equiv (\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2,$$

so that

$$\begin{aligned}
&\frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}_{\widehat{\mathcal{L}}_n}[|W_{in}|^{2+\delta} \mid X, Y, Z] \\
&\quad = \frac{1}{(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^{2+\delta}} \cdot \frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}\left[\left|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)\right|^{2+\delta} \mid X, Z\right].
\end{aligned}$$

This quantity converges to zero in probability due to the nondegeneracy condition (NDG1) and the Lyapunov condition (Lyap-1). Hence, the conditional CLT gives the desired conditional convergence (12).

Assumption 2. We begin by decomposing $T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z)$:

$$\begin{aligned}
T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z) &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i)) && (66) \\
&\quad - \frac{1}{\sqrt{n}} \sum_{i=1}^n (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(\widehat{\mu}_{n,y}(Z_i) - \mu_{n,y}(Z_i)) && (67) \\
&\equiv I_n - J_n. && (68)
\end{aligned}$$

We claim that $J_n \xrightarrow{p} 0$. Indeed, from $\mathbb{E}[J_n \mid X, Y, Z] = 0$ and the assumption $\widehat{E}'_{n,y} \xrightarrow{p} 0$ it follows

$$\mathbb{E}[J_n^2 \mid X, Y, Z] = \mathrm{Var}[J_n \mid X, Y, Z] \equiv \widehat{E}'_{n,y} \xrightarrow{p} 0.$$

Hence by Lemma 2 we have $J_n^2 \xrightarrow{p} 0$, so that $J_n \xrightarrow{p} 0$, as claimed. Next, we claim that an appropriately rescaled I_n converges conditionally to $N(0, 1)$. To this end, we apply the conditional CLT (Theorem 8) with

$$W_{in} \equiv \frac{1}{\sqrt{n}}(\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \mu_{n,y}(Z_i)) \tag{69}$$

and $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. To verify the assumptions of the conditional CLT, note first that W_{in} are independent conditionally on $\mathcal{F}_n$ by construction and satisfy $0 < \mathrm{Var}[W_{in} \mid \mathcal{F}_n] < \infty$ by the nondegeneracy assumption (NDG2). Next, observe that

$$s_n^2 \equiv \sum_{i=1}^n \mathrm{Var}[W_{in} \mid X, Y, Z] = \frac{1}{n} \sum_{i=1}^n \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i](Y_i - \mu_{n,y}(Z_i))^2 \equiv (S_n^{\widehat{\mathrm{dCRT}}})^2,$$

so that

$$\begin{aligned}
&\frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}[|W_{in}|^{2+\delta} \mid X, Y, Z] \\
&\quad = \frac{1}{(S_n^{\widehat{\mathrm{dCRT}}})^{2+\delta}} \cdot \frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |Y_i - \mu_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z].
\end{aligned}$$

Since the first factor is stochastically bounded (conclusion 99 from Lemma 7), it suffices to show that the second factor converges to zero in probability. To this end, by Lemma 2 it suffices to note that $\mathcal{L}_n \in \mathscr{L}_n^0$ and the Lyapunov assumption (Lyap-2) give

$$\begin{aligned}
&\mathbb{E}\left[\left.\frac{1}{n^{1+\delta/2}} \sum_{i=1}^n |Y_i - \mu_{n,y}(Z_i)|^{2+\delta} \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z]\right| X, Z\right] \\
&\quad = \frac{1}{n^{1+\delta/2}} \sum_{i=1}^n \mathbb{E}_{\mathcal{L}_n}[|Y_i - \mu_{n,y}(Z_i)|^{2+\delta}|Z_i] \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z] \\
&\quad \xrightarrow{p} 0.
\end{aligned} \tag{70}$$

Therefore, we may apply the conditional CLT to obtain that

$$\frac{1}{S_n^{\widehat{\mathrm{dCRT}}}} I_n \mid X, Y, Z \xrightarrow{d,p} N(0, 1).$$

Furthermore, equation (98) from Lemma 7 gives $S_n^{\widehat{\mathrm{dCRT}}} / \widehat{S}_n^{\widehat{\mathrm{dCRT}}} \xrightarrow{p} 1$, so by conditional Slutsky's theorem (Theorem 6) we conclude that

$$\frac{1}{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}} T_n^{\widehat{\mathrm{dCRT}}}(\widetilde{X}, X, Y, Z) = \frac{S_n^{\widehat{\mathrm{dCRT}}}}{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}} \left.\left(\frac{I_n}{S_n^{\widehat{\mathrm{dCRT}}}} - \frac{J_n}{S_n^{\widehat{\mathrm{dCRT}}}}\right)\right| X, Y, Z \xrightarrow{d,p} N(0, 1), \tag{71}$$

as desired. □

Corollary 7. *Under the assumptions of Theorem 9, if the non-accumulation condition* (15) *is satisfied then the* $\widehat{\mathrm{dCRT}}$ *is asymptotically equivalent to the* $\widehat{\mathrm{MX}(2)}$ *F-test.*

Proof. The desired equivalence is given by Lemma 3 with $T_n(X, Y, Z) \equiv (\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^{-1} T_n^{\widehat{\mathrm{dCRT}}}$ and $C_n(X, Y, Z) \equiv C_n^{\widehat{\mathrm{dCRT}}}(X, Y, Z)$, since ϕ_n^1 and ϕ_n^2 in the lemma statement reduce to $\widehat{\mathrm{dCRT}}$ and the $\widehat{\mathrm{MX}(2)}$ F-test, respectively. This lemma is applicable because the convergence of the critical value (76) is given by Theorem 9 and the non-accumulation condition (77) is assumed. □

C.2 Auxiliary lemmas

Lemma 2. *Let W_n be a sequence of nonnegative random variables and let $\mathcal{F}_n$ be a sequence of σ-algebras. If $\mathbb{E}[W_n \mid \mathcal{F}_n] \xrightarrow{p} 0$, then $W_n \xrightarrow{p} 0$.*

Proof. For any $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}[W_n \geq \epsilon] &= \mathbb{P}[W_n \wedge \epsilon \geq \epsilon] && (72)\\
&\leq \epsilon^{-1}\mathbb{E}[W_n \wedge \epsilon] && (73)\\
&= \epsilon^{-1}\mathbb{E}[\mathbb{E}[W_n \wedge \epsilon \mid \mathcal{F}_n]] && (74)\\
&\leq \epsilon^{-1}\mathbb{E}[\mathbb{E}[W_n \mid \mathcal{F}_n] \wedge \epsilon] \to 0, && (75)
\end{aligned}$$

where the last convergence is due to bounded convergence theorem. □

Lemma 3 (Asymptotic equivalence of tests)**.** *Consider two hypothesis tests based on the same test statistic $T_n(X,Y,Z)$ but different critical values:*

$$\phi_n^1(X,Y,Z) \equiv \mathbb{1}(T_n(X,Y,Z) > C_n(X,Y,Z)); \quad \phi_n^2(X,Y,Z) \equiv \mathbb{1}(T_n(X,Y,Z) > z_{1-\alpha}).$$

If the critical value of the first converges in probability to that of the second:

$$C_n(X,Y,Z) \xrightarrow{p} z_{1-\alpha} \tag{76}$$

and the test statistic does not accumulate near the critical value:

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\ \mathbb{P}_{\mathcal{L}_n}[|T_n(X,Y,Z) - z_{1-\alpha}| \leq \delta] = 0, \tag{77}$$

then the two tests are asymptotically equivalent:

$$\lim_{n\to\infty}\mathbb{P}[\phi_n^1(X,Y,Z) = \phi_n^2(X,Y,Z)] = 1. \tag{78}$$

Proof. Note that for any $\delta > 0$, we have

$$\begin{aligned}
&\mathbb{P}_{\mathcal{L}_n}\left[\phi_n^1(X,Y,Z) \neq \phi_n^2(X,Y,Z)\right]\\
&\quad= \mathbb{P}_{\mathcal{L}_n}\left[\min\left(z_{1-\alpha}, C_n\right) < T_n \leq \max\left(z_{1-\alpha}, C_n\right)\right]\\
&\quad= \mathbb{P}_{\mathcal{L}_n}\left[\min\left(z_{1-\alpha}, C_n\right) < T_n \leq \max\left(z_{1-\alpha}, C_n\right), |C_n - z_{1-\alpha}| \leq \delta\right]\\
&\qquad+ \mathbb{P}_{\mathcal{L}_n}\left[\min\left(z_{1-\alpha}, C_n\right) < T_n \leq \max\left(z_{1-\alpha}, C_n\right), |C_n - z_{1-\alpha}| > \delta\right]\\
&\quad\leq \mathbb{P}_{\mathcal{L}_n}\left[|T_n - z_{1-\alpha}| \leq \delta\right] + \mathbb{P}_{\mathcal{L}_n}\left[|C_n - z_{1-\alpha}| > \delta\right].
\end{aligned} \tag{79}$$

To justify the last step, suppose without loss of generality that $z_{1-\alpha} \leq C_n$. Then note that if $z_{1-\alpha} < T_n \leq C_n$ and $C_n - z_{1-\alpha} \leq \delta$ then

$$|T_n - z_{1-\alpha}| = T_n - z_{1-\alpha} \leq C_n - z_{1-\alpha} \leq \delta.$$

Taking a limsup on both sides in equation (79) and using the assumed convergence (76), we find that

$$\begin{aligned}
&\limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}\left[\phi_n^1(X,Y,Z) \neq \phi_n^2(X,Y,Z)\right] \\
&\qquad \leq \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}\left[|T_n(X,Y,Z) - z_{1-\alpha}| \leq \delta\right] + \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}\left[|C_n(X,Y,Z) - z_{1-\alpha}| > \delta\right] \\
&\qquad = \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n}\left[|T_n(X,Y,Z) - z_{1-\alpha}| \leq \delta\right].
\end{aligned}$$

Letting $\delta \to 0$ and using our assumption (77), we arrive at the claimed asymptotic equivalence. This completes the proof. □

D Proofs for Section 4 and Appendix A

For the sake of this section, we define

$$s_n^2 \equiv \mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y}|\boldsymbol{Z}]]. \tag{80}$$

D.1 Proofs of main results

Proof of Theorem 2. To show the asymptotic equivalence of variance estimates (26) it suffices to show that

$$(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2 - s_n^2 \xrightarrow{p} 0; \quad (\widehat{S}_n^{\mathrm{GCM}})^2 - s_n^2 \xrightarrow{p} 0; \quad \inf_n s_n^2 > 0. \tag{81}$$

The first of these statements is given by equation (90) in Lemma 6, the second follows from the proof of Theorem 6 in Shah and Peters (2020), and the third is a consequence of assumption (SP2) and conditional independence.

Given the asymptotic equivalence of the variance estimates (26), we can show using Lemma 3 that the GCM test is asymptotically equivalent to the $\widehat{\mathrm{MX}(2)}$ F-test. Indeed, set

$$T_n \equiv T_n^{\mathrm{GCM}} \quad \text{and} \quad C_n \equiv \frac{\widehat{S}_n^{\widehat{\mathrm{dCRT}}}}{\widehat{S}_n^{\mathrm{GCM}}} z_{1-\alpha} \tag{82}$$

Then, ϕ_n^1 is the GCM test and ϕ_n^2 is the $\widehat{\mathrm{MX}(2)}$ F-test. The asymptotic equivalence of the variance estimates (26) then implies the critical value convergence assumption (76) of Lemma 3. The non-accumulation assumption (77) is a consequence of the fact that, under the assumptions of Theorem 2, we have $T_n^{\mathrm{GCM}} \xrightarrow{d} N(0,1)$ (Shah and Peters, 2020). On the other hand, by Corollary 7 (whose conclusion holds under the assumptions of Theorem 2), we also know that $\widehat{\mathrm{dCRT}}$ is asymptotically equivalent to the $\widehat{\mathrm{MX}(2)}$ F-test. Hence, both the GCM test and the $\widehat{\mathrm{dCRT}}$ are asymptotically equivalent to the $\widehat{\mathrm{MX}(2)}$ F-test, so they are asymptotically equivalent to each other as well. □

Proof of Proposition 1. We treat the two cases separately.

Case 1. We have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}((X_i - \widehat{\mu}_{n,x}(Z_i))^2 - \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i])\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad = (s_n^2 + o_p(1)) - (s_n^2 + o_p(1)) \\
&\quad = o_p(1),
\end{aligned}
$$

where the third line follows from the convergences (87) and (84) from Lemma 6. This shows the variance consistency property (25).

Case 2. We need to show that

$$
\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(X_i)))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \xrightarrow{p} 0.
$$

By the Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i)))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad \leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]}
\end{aligned}
\tag{83}
$$

Given the assumption that $\sup_n \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty$ for some $\delta > 0$, Jensen's inequality gives

$$
\sup_n \ \mathbb{E}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]^{1+\delta/2}] \leq \sup_n \ \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}] < \infty.
$$

Therefore, the weak law of large numbers (Corollary 6) gives

$$
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \mathbb{E}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \xrightarrow{p} 0.
$$

Furthermore,

$$
\mathbb{E}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \leq \sup_n \ \mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta}]^{\frac{2}{2+\delta}} < \infty,
$$

so

$$
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] = O_p(1).
$$

On the other hand, we know $\mathrm{supp}(\mu_{n,x}(Z_i)) \subseteq \mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X}))$ for every i and n and by assumption $\widehat{\mu}_{n,x}(\boldsymbol{Z}) \subseteq \mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$ almost surely. Together with the fact that f is Lipschitz (say with Lipschitz constant L) on $\cup_{n=1}^{\infty}\mathrm{Conv}(\mathrm{supp}(\mathcal{L}_n(\boldsymbol{X})))$, it follows that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(f(\widehat{\mu}_{n,x}(Z_i)) - f(\mu_{n,x}(Z_i)))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad \leq \frac{L^2}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2 \mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] = L^2(E'_{n,x})^2 = o_p(1).
\end{aligned}$$

Combining the last two displays with equation (83) gives us the desired result. $\square$

Proof of Corollary 2. Define the event

$$A_n \equiv \{\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)\}$$

The conclusion (27) of Theorem 2 implies that $\mathbb{P}_{\mathcal{L}_n}[A_n] \to 0$, so by contiguity of $\mathcal{L}'_n$ to $\mathcal{L}_n$ we also have $\mathbb{P}_{\mathcal{L}'_n}[A_n] \to 0$. This shows the desired asymptotic equivalence of tests (28). To show the asymptotic equivalence of powers, we derive

$$\begin{aligned}
&|\mathbb{E}_{\mathcal{L}'_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) - \phi_n^{\mathrm{GCM}}(X,Y,Z)]| \\
&\qquad \leq \mathbb{E}_{\mathcal{L}'_n}[|\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) - \phi_n^{\mathrm{GCM}}(X,Y,Z)|] \\
&\qquad = \mathbb{E}_{\mathcal{L}'_n}[|\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) - \phi_n^{\mathrm{GCM}}(X,Y,Z)|\mathbb{1}(\phi_n^{\widehat{\mathrm{dCRT}}} \neq \phi_n^{\mathrm{GCM}})] \\
&\qquad \leq \mathbb{E}_{\mathcal{L}'_n}[\mathbb{1}(\phi_n^{\widehat{\mathrm{dCRT}}} \neq \phi_n^{\mathrm{GCM}})] \\
&\qquad = \mathbb{P}_{\mathcal{L}'_n}[\phi_n^{\widehat{\mathrm{dCRT}}} \neq \phi_n^{\mathrm{GCM}}] \to 0.
\end{aligned}$$

This completes the proof. $\square$

Proof of Corollary 3. Fix $\epsilon > 0$, and for each n let $\mathcal{L}_n^* \in \mathscr{L}_n^0 \cap \mathscr{R}_n$ be such that

$$\begin{aligned}
&\mathbb{P}_{\mathcal{L}_n^*}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)] \\
&\quad \geq \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z) \neq \phi_n^{\mathrm{GCM}}(X,Y,Z)] - \epsilon.
\end{aligned}$$

Applying Theorem 2 to the sequence $\mathcal{L}_n^*$ and using the asymptotic Type-I error control of the GCM test, we obtain

$$\begin{aligned}
&\limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z)] \\
&\quad \le \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} |\mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\text{GCM}}(X,Y,Z)]| + \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\text{GCM}}(X,Y,Z)] \\
&\quad \le \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} |\mathbb{E}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z)] - \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\text{GCM}}(X,Y,Z)]| \\
&\qquad + \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\text{GCM}}(X,Y,Z)] \\
&\quad \le \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)] \\
&\qquad + \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{E}_{\mathcal{L}_n}[\phi_n^{\text{GCM}}(X,Y,Z)] \\
&\quad \le \limsup_{n\to\infty} \sup_{\mathcal{L}_n \in \mathscr{L}_n^0 \cap \mathscr{R}_n} \mathbb{P}_{\mathcal{L}_n}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)] + \alpha \\
&\quad \le \limsup_{n\to\infty} \mathbb{P}_{\mathcal{L}_n^*}[\phi_n^{\widehat{\text{dCRT}}}(X,Y,Z) \neq \phi_n^{\text{GCM}}(X,Y,Z)] + \epsilon + \alpha \\
&\quad = \epsilon + \alpha.
\end{aligned}$$

Sending $\epsilon \to 0$ gives the desired conclusion. □

Proof of Theorem 4. We have

$$T_n^{\widehat{\text{ndCRT}}}(\widetilde{X},X,Y,Z) \equiv \frac{T_n^{\text{dCRT}}(\widetilde{X},X,Y,Z)}{S_n^{\text{GCM}}(\widetilde{X},X,Y,Z)} \equiv \frac{S_n^{\widehat{\text{dCRT}}}(X,Y,Z)}{S_n^{\text{GCM}}(\widetilde{X},X,Y,Z)} \cdot \frac{T_n^{\text{dCRT}}(\widetilde{X},X,Y,Z)}{S_n^{\widehat{\text{dCRT}}}(X,Y,Z)}$$

The first factor converges to 1 in probability (Lemma 8), whereas the second factor converges conditionally on X, Y, Z to $N(0,1)$ (Theorem 1). Putting these two statements together with conditional Slutsky's theorem (Theorem 6), we arrive at the convergence (49). Since the standard normal has continuous CDF we can use Lemma 1 to conclude the convergence of the critical value (50). The equivalence statement (51) follows from the convergence (50) and Lemma 3 applied with $T_n = T_n^{\text{GCM}}$ and $C_n = C_n^{\widehat{\text{ndCRT}}}$. □

Proof of Corollary 5. The proof of this corollary is directly analogous to that of Corollary 3, so we omit it for the sake of brevity.

□

D.2 Auxiliary lemmas

Lemma 4 (Conditional Jensen inequality, Davidson, 2003, Theorem 10.18)**.** *Let W be a random variable and let ϕ be a convex function, such that W and $\phi(W)$ are integrable. For any σ-algebra $\mathcal{F}$, we have the inequality*

$$\phi(\mathbb{E}[W \mid \mathcal{F}]) \le \mathbb{E}[\phi(W) \mid \mathcal{F}] \quad \textit{almost surely.}$$

Lemma 5. *Let W_n be a sequence of random variables and $\mathcal{F}_n$ a sequence of σ-algebras. If $W_n \mid \mathcal{F}_n \xrightarrow{p,p} 0$, then $W_n \xrightarrow{p} 0$.*

Proof. Let $\epsilon > 0$. Because of the assumed conditional convergence in probability, we have

$$\mathbb{P}[|W_n| > \epsilon \mid \mathcal{F}_n] \xrightarrow{p} 0.$$

By the bounded convergence theorem, it follows that

$$\mathbb{P}[|W_n| > \epsilon] = \mathbb{E}[\mathbb{P}[|W_n| > \epsilon \mid \mathcal{F}_n]] \to 0,$$

from which the conclusion follows. $\square$

Lemma 6. *Consider a sequence of laws $\mathcal{L}_n \in \mathscr{L}_n^0$. Given assumption* (SP2)*, we have*

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0; \tag{84}$$

$$\frac{1}{n}\sum_{i=1}^n (X_i - \mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0; \tag{85}$$

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i](Y_i - \mu_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0. \tag{86}$$

Given additionally assumption (SP1)*, we also have*

$$\frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - s_n^2 \xrightarrow{p} 0; \tag{87}$$

$$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0. \tag{88}$$

Given additionally the conditional Lyapunov condition (Lyap-2) *and the variance consistency condition* (25)*, we also have*

$$(S_n^{\widehat{\mathrm{dCRT}}})^2 - s_n^2 \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \mu_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0. \tag{89}$$

Given additionally the assumption $\widehat{E}'_{n,y} \xrightarrow{p} 0$, we also have

$$(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2 - s_n^2 \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i - \widehat{\mu}_{n,y}(Z_i))^2 - s_n^2 \xrightarrow{p} 0. \tag{90}$$

Proof. We prove the convergence statements in order.

Proofs of statements (84)**,** (85)**,** (86)**.** These statements are consequences of the weak law of large numbers (Corollary 6). To verify the statement (84), we note that

$$\mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]] = \mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]] \equiv s_n^2,$$

and

$$\begin{aligned}
&\sup_{i,n}\ \mathbb{E}_{\mathcal{L}_n}[|\mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]|^{1+\delta/2}] \\
&\quad = \sup_{n}\ \mathbb{E}_{\mathcal{L}_n}[|\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]|^{1+\delta/2}] \\
&\quad \leq \sup_{n}\ \mathbb{E}_{\mathcal{L}_n}[\mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z})|^{2+\delta} \mid \boldsymbol{Z}]\mathbb{E}_{\mathcal{L}_n}[|\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z})|^{2+\delta} \mid \boldsymbol{Z}]] \\
&\quad = \sup_{n}\ \mathbb{E}_{\mathcal{L}_n}[|(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))|^{2+\delta}] \\
&\quad < c_2 < \infty.
\end{aligned} \tag{91}$$

The inequality in the third line follows from the conditional Jensen inequality (Lemma 4), the equality in the fourth line follows from the conditional independence assumption, and the inequality in the fifth line follows from the $2+\delta$ moment assumption (SP2). Hence we have verified the sufficient condition (60) for the WLLN, so the convergence (84) follows. Statements (85) and (86) can be verified with similar arguments.

Proofs of statements (87) **and** (88)**.** We prove only the first, as the second will follow by symmetry. Given the convergence (85), the statement (87) will follow if we show that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu}_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] - \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad\quad - \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad \equiv (E'_{n,x})^2 - \frac{2}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i) - \mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] \\
&\quad \xrightarrow{p} 0.
\end{aligned} \tag{92}$$

The convergence $E'_{n,x} \xrightarrow{p} 0$ is assumed (SP1). To verify the convergence of the second term, note first that conditional Hölder (Lemma 12) and the derivation (91) give

$$\sup_{n}\ s_n^2 \leq \sup_{n}\ \{\mathbb{E}_{\mathcal{L}_n}[|\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{X} \mid \boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]|^{1+\delta/2}]\}^{2/(2+\delta)} < \infty, \tag{93}$$

which combined with the convergence (85) implies that

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i] = O_p(1). \tag{94}$$

Therefore, by the Cauchy-Schwartz inequality we find that

$$
\begin{aligned}
&\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{n,x}(Z_i))(\widehat{\mu}_{n,x}(Z_i)-\mu_{n,x}(Z_i))\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)^2 \\
&\quad\leq \left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)\left(\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,x}(Z_i)-\mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right) \\
&\quad\equiv \left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_{n,x}(Z_i))^2\mathrm{Var}_{\mathcal{L}_n}[Y_i \mid Z_i]\right)\cdot(E'_{n,x})^2 \\
&\quad= O_p(1)\cdot o_p(1) = o_p(1).
\end{aligned}
$$

This proves the convergence (92), which in turn implies the claimed convergence (87).

Proof of statement (89)**.** Note that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i-\mu_{n,y}(Z_i))^2 - s_n^2 \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i-\mu_{n,y}(Z_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i] \\
&\qquad+ \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i] - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i] \\
&\qquad+ \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathcal{L}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i] - s_n^2 \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i-\mu_{n,y}(Z_i))^2 - \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i] + o_p(1),
\end{aligned}
$$

where we used the variance consistency assumption (25) and the convergence result (84) to obtain the last line. Hence, it suffices to show that

$$
\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i-\mu_{n,y}(Z_i))^2 - \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i]\mathrm{Var}[Y_i \mid Z_i]\right) \xrightarrow{p} 0. \tag{95}
$$

To this end, we apply the conditional WLLN (Theorem 7) with $\mathcal{F}_n = \sigma(X, Z)$ and

$$
W_{in} = \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i \mid Z_i](Y_i-\mu_{n,y}(Z_i))^2.
$$

We check the required $1+\delta$ moment condition (55):

$$
\begin{aligned}
&\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}[|W_{in}|^{1+\delta/2}\mid\mathcal{F}_n]\\
&\equiv\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}\left[\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]^{1+\delta/2}\cdot|Y_i-\mu_{n,y}(Z_i)|^{2+\delta}\mid X,Z\right]\\
&\leq\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\mathcal{L}_n}\left[\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i-\widehat{\mu}_{n,x}(Z_i)|^{2+\delta}\mid X,Z]\cdot|Y_i-\mu_{n,y}(Z_i)|^{2+\delta}\mid X,Z\right]\\
&=\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n}\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_i-\widehat{\mu}_{n,x}(Z_i)|^{2+\delta}\mid X,Z]\cdot\mathbb{E}_{\mathcal{L}_n}\left[|Y_i-\mu_{n,y}(Z_i)|^{2+\delta}\mid Z_i\right]\\
&\xrightarrow{p}0.
\end{aligned}
$$

The inequality in the third line follows from the conditional Jensen inequality (Lemma 4), the equality in the fourth line from the assumed conditional independence, and the convergence in the fifth line from the conditional Lyaponov assumption (Lyap-2). Therefore, the conditional WLLN gives

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i]\mathrm{Var}_{\mathcal{L}_n}[Y_i\mid Z_i]\right)\\
&\quad=\frac{1}{n}\sum_{i=1}^{n}\left(\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2-\mathbb{E}_{\mathcal{L}_n}[\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2|X,Z]\right)\\
&\quad\xrightarrow{p,p}0,
\end{aligned}
$$

where the equality follows from the assumed conditional independence. Since conditional convergence in probability implies unconditional convergence in probability (Lemma 5), this verifies the claimed convergence statement (95) and completes the proof of the statement (89).

Proof of statement (90). Given the convergence (89), it suffices to show that

$$
\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\widehat{\mu}_{n,y}(Z_i))^2-\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i\mid Z_i](Y_i-\mu_{n,y}(Z_i))^2\xrightarrow{p}0 \tag{96}
$$

Given the assumption that $\widehat{E}'_{n,y}\xrightarrow{p}0$, this statement's proof is analogous to that of statement (92), so we omit it for the sake of brevity. This completes the proof of the lemma. □

Lemma 7. *Define*

$$
(S_n^{\widehat{\mathrm{dCRT}}})^2\equiv\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_i|Z_i](Y_i-\mu_{n,y}(Z_i))^2. \tag{97}
$$

Under the assumptions of Theorem 2, we have

$$\frac{(S_n^{\widehat{\mathrm{dCRT}}})^2}{(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2} \xrightarrow{p} 1 \tag{98}$$

and

$$\mathbb{P}[(S_n^{\widehat{\mathrm{dCRT}}})^2 > \epsilon] \to 1 \quad \textit{for some } \epsilon > 0. \tag{99}$$

Proof. The equivalence of variances (98) follows from the convergences (89) and (90) (Lemma 6), as well as the observation that conditional independence and the assumption (SP2) imply

$$\inf_n s_n^2 = \inf_n \ \mathbb{E}_{\mathcal{L}_n}[(\boldsymbol{X} - \mu_{n,x}(\boldsymbol{Z}))^2(\boldsymbol{Y} - \mu_{n,y}(\boldsymbol{Z}))^2] > 0. \tag{100}$$

The stochastic boundedness from below (99) follows from the latter fact and the convergence (89). $\square$

Lemma 8. *If the nondegeneracy conditions* (NDG1) *and* (NDG2) *and the conditional Lyapunov condition* (Lyap-1) *hold, then the variance estimates* $(\widehat{S}_n^{\mathrm{GCM}})^2$ *and* $(\widehat{S}_n^{\widehat{\mathrm{dCRT}}})^2$ *are equivalent under resampling:*

$$\frac{(\widehat{S}_n^{\widehat{\mathrm{dCRT}}}(X,Y,Z))^2}{(\widehat{S}_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z))^2} \xrightarrow{p} 1. \tag{101}$$

Proof. Define

$$W_{in} \equiv (\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i))(Y_i - \widehat{\mu}_{n,y}(Z_i)),$$

so that

$$(\widehat{S}_n^{\mathrm{GCM}}(\widetilde{X},X,Y,Z))^2 \equiv \frac{1}{n}\sum_{i=1}^n W_{in}^2 - \left(\frac{1}{n}\sum_{i=1}^n W_{in}\right)^2.$$

First we claim that $\frac{1}{n}\sum_{i=1}^n W_{in} \xrightarrow{p} 0$. We will use conditional WLLN (Theorem 7) with $\mathcal{F}_n \equiv \sigma(X,Y,Z)$. First note that $\mathbb{E}[W_{in} \mid \mathcal{F}_n] = 0$ by construction. We also check the moment condition (55):

$$\begin{aligned}&\frac{1}{n^{2+\delta}}\sum_{i=1}^n \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n] \\ &\qquad = \frac{1}{n^{1+\delta/2}}\frac{1}{n^{1+\delta/2}}\sum_{i=1}^n |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta}\mathbb{E}\left[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z\right] \xrightarrow{p} 0,\end{aligned}$$

where the latter convergence is by the assumption (Lyap-1). Hence we have that $\frac{1}{n}\sum_{i=1}^n W_{in} \mid \mathcal{F}_n \xrightarrow{p,p} 0$ which by Lemma 5 implies $\frac{1}{n}\sum_{i=1}^n W_{in} \xrightarrow{p} 0$.

Next we show that $\frac{1}{n}\sum_{i=1}^n W_{in}^2 - (\widehat{S_n^{\text{dCRT}}})^2 \xrightarrow{p} 0$. We will use conditional WLLN with $\mathcal{F}_n = \sigma(X, Y, Z)$, observing that $\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^n W_{in}^2 \mid \mathcal{F}_n\right) = (\widehat{S_n^{\text{dCRT}}})^2$. Next we verify the moment condition (55):

$$
\begin{aligned}
&\frac{1}{n^{1+\delta/2}}\sum_{i=1}^n \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n] && (102)\\
&\quad = \frac{1}{n^{1+\delta/2}}\sum_{i=1}^n |Y_i - \widehat{\mu}_{n,y}(Z_i)|^{2+\delta}\mathbb{E}\left[|\widetilde{X}_i - \widehat{\mu}_{n,x}(Z_i)|^{2+\delta} \mid X, Z\right] \xrightarrow{p} 0, && (103)
\end{aligned}
$$

where the latter convergence is by the assumption (Lyap-1). Hence we have that $\frac{1}{n}\sum_{i=1}^n W_{in}^2 - (\widehat{S_n^{\text{dCRT}}})^2 \mid \mathcal{F}_n \xrightarrow{p,p} 0$ which by Lemma 5 implies that $\frac{1}{n}\sum_{i=1}^n W_{in}^2 - (\widehat{S_n^{\text{dCRT}}})^2 \xrightarrow{p} 0$.

Combining both of these results we find that

$$
(\widehat{S_n^{\text{ndCRT}}}(\widetilde{X}, X, Y, Z))^2 - (\widehat{S_n^{\text{dCRT}}}(X, Y, Z))^2 \xrightarrow{p} 0.
$$

Now using the nondegeneracy condition (NDG1) we can conclude that (101) holds true, as desired. $\square$

E Proofs for Section 5

The goal of this section is to prove our main optimality result (Theorem 3) and Corollary 4. The idea of the proof of Theorem 3 is to reduce the problem to a semiparametric testing problem, and then to use existing semiparametric optimality theory. To this end, we first review the relevant semiparametric theory (Section E.1). Then we leverage this theory to prove Theorem 3 (Section E.2) and verify Corollary 4 (Section E.3). Finally, we carry out deferred semiparametric computations (Section E.4).

E.1 Semiparametric preliminaries

Consider a semiparametric model parameterized by

$$
(\beta, g) \in \mathbb{R} \times \mathcal{H}_g \subseteq \mathbb{R} \times L^2(\nu), \tag{104}
$$

where ν is a measure on $\mathbb{R}^p$ and $\mathcal{H}_g \subseteq L^2(\nu)$ is a linear subspace. First, we define a notion of local Type-I error control within the context of the semiparametric model.

Definition 4. Fix a point $g_0 \in \mathcal{H}_g$, and define $\theta_0 \equiv (0, g_0)$. A sequence of tests ϕ_n of $H_0 : \beta = 0$ has asymptotic Type-I error control at θ_0 relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ if, for each submodel $t \mapsto \mathcal{L}_{(0,g_t)}$ with score in $\dot{\mathcal{L}}_{\theta_0}$ along which β is differentiable, we have

$$
\limsup_{n\to\infty} \ \mathbb{E}_{\mathcal{L}_{(0,g_{1/\sqrt{n}})}}[\phi_n(W)] \leq \alpha. \tag{105}
$$

This definition is most similar to that of Choi, Hall, and Schick (1996), except the latter paper does not explicitly use the language of tangent spaces; our definition accommodates Type-I error control over more restricted sets of null distributions reflecting regularity conditions. Next we state a version of the classic semiparametric optimality result:

Theorem 10 (Theorem 1 in Choi, Hall, and Schick, 1996, Theorem 25.44 in Van Der Vaart, 1998, Theorem 18.12 in Kosorok, 2008)**.** *Consider a semiparametric model* $\{\mathcal{L}_{\beta,g} : (\beta, g) \in \mathbb{R} \times \mathcal{H}_g\}$ *and a point* $\theta_0 \equiv (0, g_0)$ *for some* $g_0 \in \mathcal{H}_g$*. Suppose* β *is differentiable at* $\mathcal{L}_{\theta_0}$ *relative to the tangent space* $\dot{\mathcal{L}}_{\theta_0}$ *with efficient influence function* $\widetilde{S}/\widetilde{I}(\theta_0)$*, where* $\widetilde{S}$ *is the efficient score and* $\widetilde{I}(\theta_0) > 0$ *is the efficient information. For any sequence of tests* ϕ_n *of* $H_0 : \beta = 0$ *with asymptotic Type-I error control at* θ_0 *relative to the tangent space* $\dot{\mathcal{L}}_{\theta_0}$ *and any differentiable submodel* $\mathcal{L}_t = \mathcal{L}_{(th_\beta, g_t)}$ *with score in* $\dot{\mathcal{L}}_{\theta_0}$ *we have*

$$\limsup_{n\to\infty} \mathbb{E}_{\mathcal{L}_{1/\sqrt{n}}}[\phi_n(W)] \leq 1 - \Phi(z_{1-\alpha} - h_\beta \cdot \widetilde{I}(\theta_0)^{1/2}). \tag{106}$$

This bound is achieved by the efficient score test $\phi_n^{\text{opt}}(X, Y, Z) \equiv \mathbb{1}(T_n^{\text{opt}}(X, Y, Z) > z_{1-\alpha})$*, where*

$$T_n^{\text{opt}}(X, Y, Z) \equiv \frac{1}{\widetilde{I}(\theta_0)^{1/2} n^{1/2}} \sum_{i=1}^{n} \widetilde{S}(X_i, Y_i, Z_i). \tag{107}$$

In other words,

$$\begin{aligned} \lim_{n\to\infty} \mathbb{E}_{\mathcal{L}_{1/\sqrt{n}}}[\phi_n^{\text{opt}}(W)] &= \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{1/\sqrt{n}}}[T_n^{\text{opt}}(X, Y, Z) > z_{1-\alpha}] \\ &= 1 - \Phi(z_{1-\alpha} - h_\beta \cdot \widetilde{I}(\theta_0)^{1/2}). \end{aligned} \tag{108}$$

This result is like Choi, Hall, and Schick (1996, Theorem 1), except it explicitly deals with tangent spaces. On the other hand, the result is like Van Der Vaart (1998, Theorem 25.44) or Kosorok (2008, Theorem 18.12), except it is written in terms of semiparametric models and assumes Type-I error control in the sense of Definition 4 above. By comparison, Van Der Vaart (1998) and Kosorok (2008) assume Type-I error control at each point $(0, g)$ for $g \in \mathcal{H}_g$. By inspection of the proof of Van Der Vaart (1998, Theorem 25.44), only local Type-I error control (Definition 4) is actually needed. In this sense, Theorem 10 can be verified using the same proof as that of Van Der Vaart (1998, Theorem 25.44), specializing to the case of semiparametric models.

E.2 Proof of Theorem 3

To apply the semiparametric theory from the previous section, the following lemma (proved in Section E.4) identifies the tangent space, the efficient score, and the efficient information at $\mathcal{L}_{\theta_0}$.

Lemma 9. [Proof] *In the context of the semiparametric model* (30)*, suppose the following assumptions hold:*

$$s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]] > 0; \quad (109)$$

$$\ddot{\psi} = K > 0 \text{ and } \mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty \text{ OR } \mathrm{supp}(\boldsymbol{X}, \boldsymbol{Z}) \text{ is compact and } \mathcal{H}_g \subseteq C(\mathbb{R}^p), \quad (110)$$

$$\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\cdot\,] \in \mathcal{H}_g. \quad (111)$$

For each $h = (h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$, the parametric submodel $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ is differentiable in quadratic mean at $t = 0$ with score function

$$S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = \big(\boldsymbol{X} h_\beta + h_g(\boldsymbol{Z})\big)(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \quad (112)$$

and satisfies the following local asymptotic normality:

$$\log \prod_{i=1}^{n} \frac{\mathrm{d}\mathcal{L}_{\theta_n(h)}}{\mathrm{d}\mathcal{L}_{\theta_0}}(X_i, Y_i, Z_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S(X_i, Y_i, Z_i) - \frac{1}{2}\mathrm{Var}_{\mathcal{L}_{\theta_0}}[S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})] + o_{\mathcal{L}_{\theta_0}}(1). \quad (113)$$

The parameter β is differentiable at $\mathcal{L}_{\theta_0}$ relative to the tangent space

$$\dot{\mathcal{L}}_{\theta_0} \equiv \{\big(\boldsymbol{X} h_\beta + h_g(\boldsymbol{Z})\big)(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) : (h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g\} \quad (114)$$

with efficient score function

$$\widetilde{S}(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) = (\boldsymbol{X} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]), \quad (115)$$

efficient information

$$\widetilde{I}(\theta_0) = s^2(\theta_0) \equiv \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]], \quad (116)$$

and efficient influence function equal to the ratio of the efficient information and the efficient score.

Note that assumptions (110) and (111) of Lemma 9 are the same as assumptions (37) and (38) of Theorem 3 in the main text; they are restated here for the reader's convenience. Using Lemma 9 in conjunction with Theorem 10, we can prove Theorem 3.

Proof of Theorem 3. Let ϕ_n be a level α test of H_0 as defined in equation (33), and fix $g_0 \in \mathcal{S}$. By assumption (39), $\theta_n(0, h_g) \in \mathscr{R}$ for all $h_g \in \mathcal{H}_g$ for all sufficiently large n. Therefore, ϕ_n also has asymptotic Type-I error control at $\theta_0 \equiv (0, g_0)$ relative to the tangent space $\dot{\mathcal{L}}_{\theta_0}$ (114) in the sense of Definition 4. Indeed, it suffices to take submodels $t \mapsto \mathcal{L}_{(0,g_t)}$ for $g_t = g_0 + th_g$ and $h_g \in \mathcal{H}_g$, so that $\mathcal{L}_{(0,g_{1/\sqrt{n}})} = \mathcal{L}_{\theta_n(0,h_g)}$. By Lemma 9 (applicable because its first assumption (109) is implied by assumption (36) of Theorem 3 and its last two assumptions are also assumed by Theorem 3), the assumptions of Theorem 10 are met with efficient score $\widetilde{S}$ (115) and efficient information $s^2(\theta_0)$ (116), so taking submodels $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ we find

$$\limsup_{n\to\infty}\ \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n(X, Y, Z)] \le 1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)). \quad (117)$$

On the other hand, because $\mathcal{L}_{\theta_0} \in \mathscr{R}$, it follows that

$$
\begin{aligned}
T_n^{\mathrm{GCM}}(X,Y,Z) &= \frac{1}{s(\theta_0)\sqrt{n}} \sum_{i=1}^{n} (X_i - \mathbb{E}_{\mathcal{L}_{\theta_0}}[X_i \mid Z_i])(Y_i - \mathbb{E}_{\mathcal{L}_{\theta_0}}[Y_i \mid Z_i]) + o_{\mathcal{L}_{\theta_0}}(1) \\
&= \frac{1}{\widetilde{I}(\theta_0)^{1/2} n^{1/2}} \sum_{i=1}^{n} \widetilde{S}(X_i, Y_i, Z_i) + o_{\mathcal{L}_{\theta_0}}(1) \\
&= T_n^{\mathrm{opt}}(X,Y,Z) + o_{\mathcal{L}_{\theta_0}}(1).
\end{aligned}
\tag{118}
$$

The first equality follows from the proof of Theorem 6 in Shah and Peters (2020), the second follows from the derivations of the efficient score (115) and efficient information (116) in Lemma 9, and the third from equation (107) in Theorem 10. From the local asymptotic normality (113) it follows that $\prod_{i=1}^n \mathcal{L}_{\theta_n(h)}$ and $\prod_{i=1}^n \mathcal{L}_{\theta_0}$ are contiguous by Le Cam's first lemma (Van Der Vaart, 1998, Example 6.5). We therefore find that

$$
1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)) = \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{\theta_n(h)}}[T_n^{\mathrm{opt}} > z_{1-\alpha}] = \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{\theta_n(h)}}[T_n^{\mathrm{GCM}} > z_{1-\alpha}].
$$

The first inequality follows from the conclusion (108) of Theorem 10 and the second equality follows from equation (118) and Le Cam's first lemma. Therefore we have shown that for any $h \in (0,\infty) \times \mathcal{H}_g$ and any level α conditional independence test ϕ_n, we have

$$
\limsup_{n\to\infty} \mathbb{E}_{\mathcal{L}_{\theta_n(h)}}[\phi_n(X,Y,Z)] \le 1 - \Phi(z_{1-\alpha} - h_\beta \cdot s(\theta_0)) = \lim_{n\to\infty} \mathbb{P}_{\mathcal{L}_{\theta_n(h)}}(T_n^{\mathrm{GCM}} > z_{1-\alpha}).
$$

This shows that ϕ_n^{GCM} is LAUMP(g_0) and verifies the claimed asymptotic power (40). Furthermore, since $g_0 \in \mathcal{S}$ was chosen arbitrarily, it follows that ϕ_n^{GCM} is also LAUMP($\mathcal{S}$). This completes the proof. $\square$

E.3 Proof of Corollary 4

It suffices to verify each of the four assumptions of Theorem 3.

Verification of assumption (36)**.** Note that assumption (SP2) is satisfied because by construction, $\boldsymbol{X} - \mu_x(\boldsymbol{Z})$ and $\boldsymbol{Y} - \mu_y(\boldsymbol{Z})$ are independent standard normal random variables for any $\mathcal{L} \in \mathscr{R}$. Next, let

$$
k(z,z') = \sum_{j=1}^{\infty} \lambda_j e_j(z) e_j(z') \tag{119}
$$

be an eigendecomposition of the Sobolev kernel k with eigenfunctions e_j orthonormal with respect to Unif$[0,1]$. To verify assumption (SP1) given assumption (SP2), it suffices to prove the following statements (Shah and Peters, 2020, Theorem 11 and Remark 12):

$$\begin{aligned} \mathrm{Var}_{\mathcal{L}}[\boldsymbol{X}|\boldsymbol{Z}], \mathrm{Var}_{\mathcal{L}}[\boldsymbol{Y}|\boldsymbol{Z}] \leq \sigma^2 < \infty \quad \text{almost surely, for all } \mathcal{L} \in \mathscr{R}; \\ \sup_{\mathcal{L}\in\mathscr{R}} \max(\|\mu_{n,x}\|_{W^{1,2}[0,1]}, \|\mu_{n,y}\|_{W^{1,2}[0,1]}) < \infty; \\ \sum_{j=1}^{\infty} \lambda_j < \infty. \end{aligned}$$

The first two of these statements follow directly from the construction of $\mathscr{R}$. The third follows from the eigendecomposition of the Sobolev kernel under the uniform measure on $[0,1]$ (Wainwright, 2019, Example 12.23) with $\lambda_j = (\frac{2}{(2j-1)\pi})^2$.

Verification of assumption (37)**.** Since we are using the normal exponential family, we have $\ddot{\psi} = 1$. Furthermore,

$$\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] = \mathbb{E}_{\mathcal{L}_{x,z}}[\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2|\boldsymbol{Z}]] = \mathbb{E}_{\mathcal{L}_{x,z}}[\mu_{0x}(\boldsymbol{Z})^2 + 1] < \infty, \tag{120}$$

since $\mu_{0x} \in W^{(1,2)}([0,1]) \subseteq L^2([0,1])$.

Verification of assumption (38)**.** By construction, $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}|\cdot\,] = \mu_{0x} \in W^{1,2}([0,1])$.

Verification of assumption (39)**.** This assumption is a consequence of the openness of the ball $B_{W^{1,2}}(0, C)$ in $W^2([0,1])$.

E.4 Proof of Lemma 9

Differentiability of parametric submodels. Consider the parametric submodel $t \mapsto \mathcal{L}_{(th_\beta, g_0+th_g)}$ for some $(h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$, and denote

$$\eta_t(x,z) \equiv xth_\beta + g_0(z) + th_g(z).$$

Letting λ_y be the base measure of the exponential family f_η, we denote $\lambda \equiv \mathcal{L}_{x,z} \times \lambda_y$ and $\mathrm{d}\mathcal{L}_{(th_\beta,g_0+th_g)}(x,y,z)/\mathrm{d}\lambda$ the density of the parametric model for $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z})$ with respect to λ. According to Van Der Vaart (1998, Lemma 7.6), this submodel is differentiable in quadratic mean at $t = 0$ if the map

$$t \mapsto \sqrt{\frac{\mathrm{d}\mathcal{L}_{(th_\beta,g_0+th_g)}}{d\lambda}(x,y,z)} = \sqrt{\frac{\mathrm{d}f_{\eta_t}}{d\lambda_y}(y)} = \exp\left(\frac{y\eta_t - \psi(\eta_t)}{2}\right)$$

is continuously differentiable at $t = 0$ for each $(x, y, z) \in \mathbb{R}^{1+1+p}$ and the elements of the Fisher information matrix are well-defined and continuous at $t = 0$. To show continuous differentiability of the square root density, we compute that

$$\frac{\partial}{\partial t}\sqrt{\frac{\mathrm{d}\mathcal{L}_{(th_\beta,g_0+th_g)}}{d\lambda}(x,y,z)} = \exp\left(\frac{y\eta_t - \psi(\eta_t)}{2}\right) \cdot \frac{(y - \dot{\psi}(\eta_t))}{2} \cdot (xh_\beta + h_g(z)).$$

The linearity of η_t in t and the smoothness of ψ imply the continuous differentiability of the above function in t.

Next consider the information matrix

$$\begin{aligned} I_t &\equiv \mathbb{E}_{\mathcal{L}_{(th_\beta, g_0+th_g)}}\left[\left(\frac{\partial}{\partial t}\log\frac{\mathrm{d}\mathcal{L}_{(th_\beta,g_0+th_g)}}{d\lambda}(\boldsymbol{X},\boldsymbol{Y},\boldsymbol{Z})\right)^2\right] \\ &= \mathbb{E}_{\mathcal{L}_{(th_\beta, g_0+th_g)}}\left[\left(\frac{\partial}{\partial t}(\boldsymbol{Y}\eta_t(\boldsymbol{X},\boldsymbol{Z}) - \psi(\eta_t(\boldsymbol{X},\boldsymbol{Z})))\right)^2\right] \\ &= \mathbb{E}_{\mathcal{L}_{(\beta_t,g_t)}}[(\boldsymbol{Y}-\dot{\psi}(\eta_t(\boldsymbol{X},\boldsymbol{Z})))^2(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2 \\ &= \mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2\ddot{\psi}(\eta_t(\boldsymbol{X},\boldsymbol{Z}))\right]. \end{aligned}$$

We must show that I_t is well-defined and continuous at $t=0$. By assumption (110), either $\ddot{\psi} = K > 0$ and $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty$ or $(\boldsymbol{X},\boldsymbol{Z})$ is compactly supported and $\mathcal{H}_g \subseteq C(\mathbb{R}^p)$. If $\ddot{\psi} = K > 0$ and $\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] < \infty$, then we have

$$\begin{aligned} I_0 = C\mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2\right] &\leq 2Kh_\beta^2\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] + 2K\mathbb{E}_{\mathcal{L}_{x,z}}[h_g^2(\boldsymbol{Z})] \\ &= 2Kh_\beta^2\mathbb{E}_{\mathcal{L}_{x,z}}[\boldsymbol{X}^2] + 2K\|h_g\|^2_{L^2(\nu)} < \infty. \end{aligned}$$

Note that, for the sake of this proof, we denote

$$\nu \equiv \mathcal{L}_{x,z}(\boldsymbol{Z}), \quad \text{so that} \quad \|h_g\|^2_{L^2(\nu)} \equiv \mathbb{E}_{\mathcal{L}_{x,z}}[h_g(\boldsymbol{Z})^2] < \infty \quad \text{for } h_g \in \mathcal{H}_g. \tag{121}$$

Hence, I_0 is well-defined. I_t is also continuous at $t=0$ because it does not depend on t.

On the other hand, suppose $(\boldsymbol{X},\boldsymbol{Z})$ is compactly supported and $\mathcal{H}_g \subseteq C(\mathbb{R}^p)$. The quantity inside the expectation defining I_0 is a bounded random variable, because the assumed continuity of h_g implies that this quantity is a continuous function of a random vector $(\boldsymbol{X},\boldsymbol{Z})$ with compact support. Hence, I_0 is well-defined because it is the expectation of a bounded random variable. To show continuity of I_t, note that by the assumed continuity of h_g and compact support of $(\boldsymbol{X},\boldsymbol{Z})$ we have $\sup_{t\leq 1}\eta_t(\boldsymbol{X},\boldsymbol{Z}) \leq B < \infty$ almost surely. Therefore, for $t \leq 1$, we have

$$\begin{aligned} |I_t - I_0| &= \left|\mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2(\ddot{\psi}(\eta_t) - \ddot{\psi}(\eta_0))\right]\right| \\ &\leq \mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2|\ddot{\psi}(\eta_t) - \ddot{\psi}(\eta_0)|\right] \\ &\leq \mathbb{E}_{\mathcal{L}_{x,z}}\left[(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))^2\sup_{|b|\leq B}|\dddot{\psi}(b)|\cdot|\eta_t - \eta_0|\right] \\ &\leq \sup_{|b|\leq B}|\dddot{\psi}(b)|\cdot\mathbb{E}_{\mathcal{L}_{x,z}}\left[|\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})|^3\right]\cdot t. \end{aligned}$$

We have $\sup_{|b|\leq B}|\dddot{\psi}(b)| < \infty$ because $\dddot{\psi}$ is a continuous function, and $\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})$ almost surely bounded as before. Therefore, we conclude that $|I_t - I_0| \to 0$ as $t \to 0$, so I_t is indeed continuous at 0.

Hence, we conclude by Van Der Vaart (1998, Lemma 7.6) that the parametric submodel $t \mapsto \mathcal{L}_{(th_\beta, g_0 + th_g)}$ is differentiable in quadratic mean at $t = 0$ with score function

$$\begin{aligned} S(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) &= \frac{\partial}{\partial t}(\boldsymbol{Y}\eta_t(\boldsymbol{X}, \boldsymbol{Z}) - \psi(\eta_t(\boldsymbol{X}, \boldsymbol{Z})))\bigg|_{t=0} \\ &= (\boldsymbol{Y} - \dot{\psi}(\eta_t(\boldsymbol{X}, \boldsymbol{Z})))(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))\Big|_{t=0} \\ &= (\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z}))(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]), \end{aligned}$$

as claimed (112).

Local asymptotic normality of parametric submodels. The local asymptotic normality of parametric submodels (113) follows from the previously established quadratic mean differentiability (Van Der Vaart, 1998, Theorem 7.2).

Efficient score and information. For $(h_\beta, h_g) \in \mathbb{R} \times \mathcal{H}_g$, define the score operator

$$\begin{aligned} A(h) &\equiv \big(\boldsymbol{X}h_\beta + h_g(\boldsymbol{Z})\big)(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \\ &= \boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_\beta + (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z}) \\ &\equiv A_\beta(h_\beta) + A_g(h_g). \end{aligned} \tag{122}$$

The tangent space $\dot{\mathcal{L}}_{\theta_0}$ (114) can then be expressed as the range of A:

$$\dot{\mathcal{L}}_{\theta_0} = A(\mathbb{R} \times \mathcal{H}_g). \tag{123}$$

As discussed in Van Der Vaart, 1998, Section 25.4, the efficient score for β is

$$\tilde{S} = A_\beta - \Pi_{\beta,g}A_\beta, \tag{124}$$

where $\Pi_{\beta,g}$ is the orthogonal projection onto the closure $\overline{A_g(\mathcal{H}_g)}$ of the nuisance tangent space $A_g(\mathcal{H}_g)$ in $L^2(\mathcal{L}_{\theta_0})$. In other words,

$$\Pi_{\beta,g}A_\beta = \underset{W \in \overline{A_g(\mathcal{H}_g)}}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})}. \tag{125}$$

To compute this projection, we first claim that the extended operator $A_g : L^2(\nu) \to L^2(\mathcal{L}_{\theta_0})$ is continuous and that $A_g^* A_g$ is continuously invertible. To verify continuity of A_g, note that for $h_g \in L^2(\nu)$ we have

$$\begin{aligned} \|A_g(h_g)\|^2_{L^2(\mathcal{L}_{\theta_0})} &= \mathbb{E}_{\mathcal{L}_{\theta_0}}\left[\Big((\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])h_g(\boldsymbol{Z})\Big)^2\right] \\ &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])^2 h_g^2(\boldsymbol{Z})] \\ &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\ddot{\psi}(g(\boldsymbol{Z}))h_g^2(\boldsymbol{Z})] \\ &\le C\mathbb{E}_{\mathcal{L}_{\theta_0}}[h_g^2(\boldsymbol{Z})] \\ &= C\|h_g\|^2_{L^2(\nu)}. \end{aligned} \tag{126}$$

Next, we derive the adjoint operator $A_g^* : L^2(\mathcal{L}_{\theta_0}) \to L^2(\nu)$. For a random variable $W \in L^2(\mathcal{L}_{\theta_0})$, we have

$$\begin{aligned}
\langle W, A_g h_g \rangle_{L^2(\mathcal{L}_{\theta_0})} &= \mathbb{E}_{\mathcal{L}_{\theta_0}} \Big[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) h_g(\boldsymbol{Z}) \Big] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \mid \boldsymbol{Z}] h_g(\boldsymbol{Z})] \\
&= \langle \mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \cdot\,]) \mid \cdot\,], h_g \rangle_{L^2(\nu)}.
\end{aligned} \tag{127}$$

It follows that

$$(A_g^* W)(z) = \mathbb{E}_{\mathcal{L}_{\theta_0}}[W(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z]. \tag{128}$$

Next we derive that

$$\begin{aligned}
(A_g^* A_g h_\eta)(z) &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[A_g h_\eta (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) h_g(\boldsymbol{Z})(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z] h_g(z) \\
&= \ddot{\psi}(g(z)) h_g(z).
\end{aligned}$$

The assumption (110) implies that $\ddot{\psi}(g_0(\boldsymbol{Z})) \geq c > 0$ almost surely, since either $\ddot{\psi}$ is a nonzero constant or $g_0(\boldsymbol{Z})$ belongs to a compact set almost surely and therefore $\ddot{\psi}(g_0(\boldsymbol{Z}))$ belongs to the range of a positive continuous function applied to a compact set. From this it follows that $S_g^* S_g$ is continuously invertible. Because A_g is a continuous linear operator with continuously invertible $A_g^* A_g$, it follows that $A_g(L^2(\nu))$ is closed and that $A_g(A_g^* A_g)^{-1} A_g^*$ is the orthogonal projection onto this space. Next let us compute the orthogonal projection of the score A_β onto $A_g(L^2(\nu))$. We have

$$\begin{aligned}
(A_g^* A_\beta)(z) &= \mathbb{E}_{\mathcal{L}_{\theta_0}}[A_\beta(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z]) \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z} = z])^2 \mid \boldsymbol{Z} = z] \\
&= \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z} = z] \ddot{\psi}(g(z)),
\end{aligned} \tag{129}$$

and therefore

$$\begin{aligned}
A_g(A_g^* A_g)^{-1} A_g^* A_\beta &= (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])((A_g^* A_g)^{-1} A_g^* A_\beta)(\boldsymbol{Z}) \\
&= (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}] \\
&= A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]).
\end{aligned} \tag{130}$$

Therefore, we have

$$\operatorname*{arg\,min}_{W \in A_g(L^2(\nu))} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]). \tag{131}$$

Since $A_g(L^2(\nu))$ is closed, it follows that $\overline{A_g(\mathcal{H}_g)} \subseteq A_g(L^2(\nu))$. Together with the assumption (111) that $\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,] \in \mathcal{H}_g$ and the definition of the effective score as a projection onto $A_g(\mathcal{H}_g)$ (125), we deduce that

$$\Pi_{\beta,g} A_\beta = \underset{W \in \overline{A_g(\mathcal{H}_g)}}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = \underset{W \in A_g(L^2(\nu))}{\arg\min} \|A_\beta - W\|_{L^2(\mathcal{L}_{\theta_0})} = A_g(\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \cdot\,]).$$

Therefore, the efficient score is

$$\begin{aligned} \tilde{S} &= A_\beta - \Pi_{\beta,g} A_\beta \\ &= \boldsymbol{X}(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]) - (\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}])\mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}] \\ &= (\boldsymbol{X} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X} \mid \boldsymbol{Z}])(\boldsymbol{Y} - \mathbb{E}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y} \mid \boldsymbol{Z}]), \end{aligned} \tag{132}$$

as claimed (115). From there we find that the efficient information is

$$\widetilde{I}_{\theta_0} = \mathrm{Var}_{\mathcal{L}_{\theta_0}}[\tilde{S}] = \mathbb{E}_{\mathcal{L}_{\theta_0}}[\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{X}|\boldsymbol{Z}]\mathrm{Var}_{\mathcal{L}_{\theta_0}}[\boldsymbol{Y}|\boldsymbol{Z}]] \equiv s^2(\theta_0), \tag{133}$$

also as claimed (116).

Differentiability of β and efficient influence function. By Van Der Vaart (1998, Lemma 25.25), the differentiability of β at $\mathcal{L}_{\theta_0}$ with respect to the tangent set $\dot{\mathcal{L}}_{\theta_0}$ follows from the quadratic mean differentiability proved above and the assumption that $\widetilde{I}_{\theta_0} = s^2(\theta_0) > 0$ (109). The same lemma gives the efficient influence function as the ratio of the efficient score and the efficient information. This completes the proof.

F Additional material related to simulations.

In this section, we present details about existing robustness simulation setups (Section F.1), investigate the trade-off between using lasso and post-lasso (Section F.2) and present the complete simulation results (Section F.3).

F.1 Simulation setup in literature

Here we provide details about the simulation setups considered in Candès et al., 2018; Liu et al., 2022; Li and Liu, 2022.

Liu et al. (2022). In this paper, the authors consider the double high-dimensional linear model. Suppose $\{Z_i, Y_i\}_{i=1}^n$ is $n = 800$ iid data and $Z_i \sim N(\mathbf{0}, \mathbf{\Sigma}_p)$ where $p = 800$ and $\mathbf{\Sigma}_p$ is chosen to be AR(1) and the autocorrelation is set to be 0.5. Then they consider $Y_i = Z_i^\top \beta + \epsilon$ where β is vector of dimension p and only $s = 50$ is set to be nonzero with magnitude $\nu = 0.175$ and random sign. They consider two ways to set the nonzero components of β: spacing these non-zero coefficients equally or choosing them to be the first 50 coefficients of β. The authors consider, rather than testing conditional independence, the false discovery rate (FDR) of variable selection.

Candès et al. (2018). In this paper, the authors consider a bit different setting where $Y|Z$ is now a high-dimensional logistic model. The sample size $n = 800$ and $Z_i \sim N(\mathbf{0}, \mathbf{\Sigma}_p)$ where $p = 1500$ and $\mathbf{\Sigma}_p$ is chosen to be AR(1) and the autocorrelation is set to be 0.3. After the sampling, the design matrix is centered and every column is normalized to have norm 1. Similarly, only $s = 50$ coordinates of β are set to be nonzero and the sign is random whereas the magnitude is set to be $\nu = 20$. They set a randomly-chosen set of $s = 50$ coefficients of β to be nonzero and consider again the FDR control.

Li and Liu (2022). In this paper, a similar setting is considered while Z is a data matrix with row $n = 250$ and column $p = 500$ where each row is sampled from $N(0, \mathbf{\Sigma}_p)$ and $\mathbf{\Sigma}_p$ is an AR(1) matrix with autocorrelation 0.5. However, a crucial difference in this paper is the way to set $\mathbb{E}(X|Z)$ and $\mathbb{E}(Y|Z)$. As for X, it is generated by a linear predictor $X = Z\gamma + \epsilon$ where γ is a p dimensional vector with first $s = 5$ is nonzero and the other coordinate remains zero and ϵ follows the standard normal distribution. The sign of each coordinate is randomly chosen and the magnitude is set to be $\nu = 0.3$. As for Y, we set $\beta = \gamma$ and $Y = Z\beta + \xi$ where ξ follows standard normal distribution such that ξ is independent of ϵ. We can see that X and Y support on the same subset of Z so that the marginal association between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is much larger than that in first two simulation designs.

F.2 Comparing the lasso and post-lasso estimation methods

Compare to the lasso estimation method for $\mu_{n,x}$ and $\mu_{n,y}$, the post-lasso estimation method results in estimates with lower bias but higher variance. This impacts the Type-I error and power of the inferential methods in different ways. For Type-I error, it is only important to have good estimates for $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}_{\mathcal{A}}] \equiv \boldsymbol{Z}_{\mathcal{A}}^T\beta_{\mathcal{A}}$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}_{\mathcal{A}}] \equiv \boldsymbol{Z}_{\mathcal{A}}^T\gamma_{\mathcal{A}}$, where $\mathcal{A} \subseteq \{1, \dots, p\}$ denotes the set of variables active in both $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Z}]$ and $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}]$. Indeed, only the shared active coordinates of $\boldsymbol{Z}$ act as confounders. On the other hand, for power, it is important to have a good estimate for the entire function $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}] = \boldsymbol{Z}^T\gamma$ (Katsevich and Ramdas, 2022). Therefore, we examine the mean-squared estimation error for $\boldsymbol{Z}_{\mathcal{A}}^T\beta_{\mathcal{A}}$ and $\boldsymbol{Z}_{\mathcal{A}}^T\gamma_{\mathcal{A}}$ in one of our null simulation settings as well as the mean-squared estimation error for $\boldsymbol{Z}^T\beta$ and $\boldsymbol{Z}^T\gamma$ in one of our alternative simulation settings (Figure 5). We find that the post-lasso does a better job estimating the shared active coefficients in the null setting, so the reduced bias in estimating these shared coefficients outweighs the increased variance. On the other hand, the lasso does a better job estimating the entire set of coefficients, so in this case the increased variance outweighs the reduced bias. This explains why the post-lasso-based methods have improved Type-I error control but worse power than the lasso-based methods.

F.3 Additional simulation results

Figures 6-13 present the complete simulation results across the null and alternative, Gaussian and binary, and supervised and unsupervised settings.

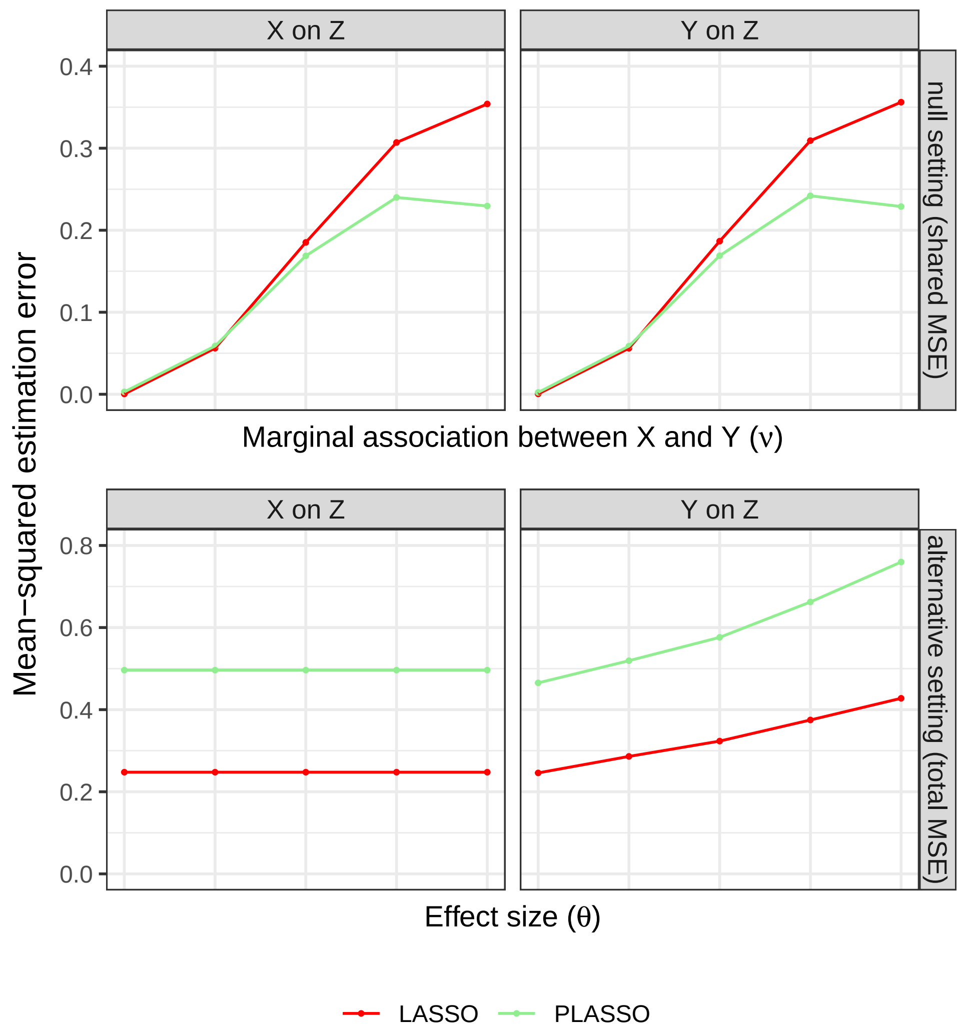

Figure 5: MSE on shared variables and total variables: first column displays the MSE of lasso and post-lasso on shared active variables $Z_{\mathcal{A}}$ and the second column displays the MSE of lasso and post-lasso on total variables. All the experiments are carried out with GCM statistic with $n = 100, d = 400, s = 5, \rho = 0.4$.

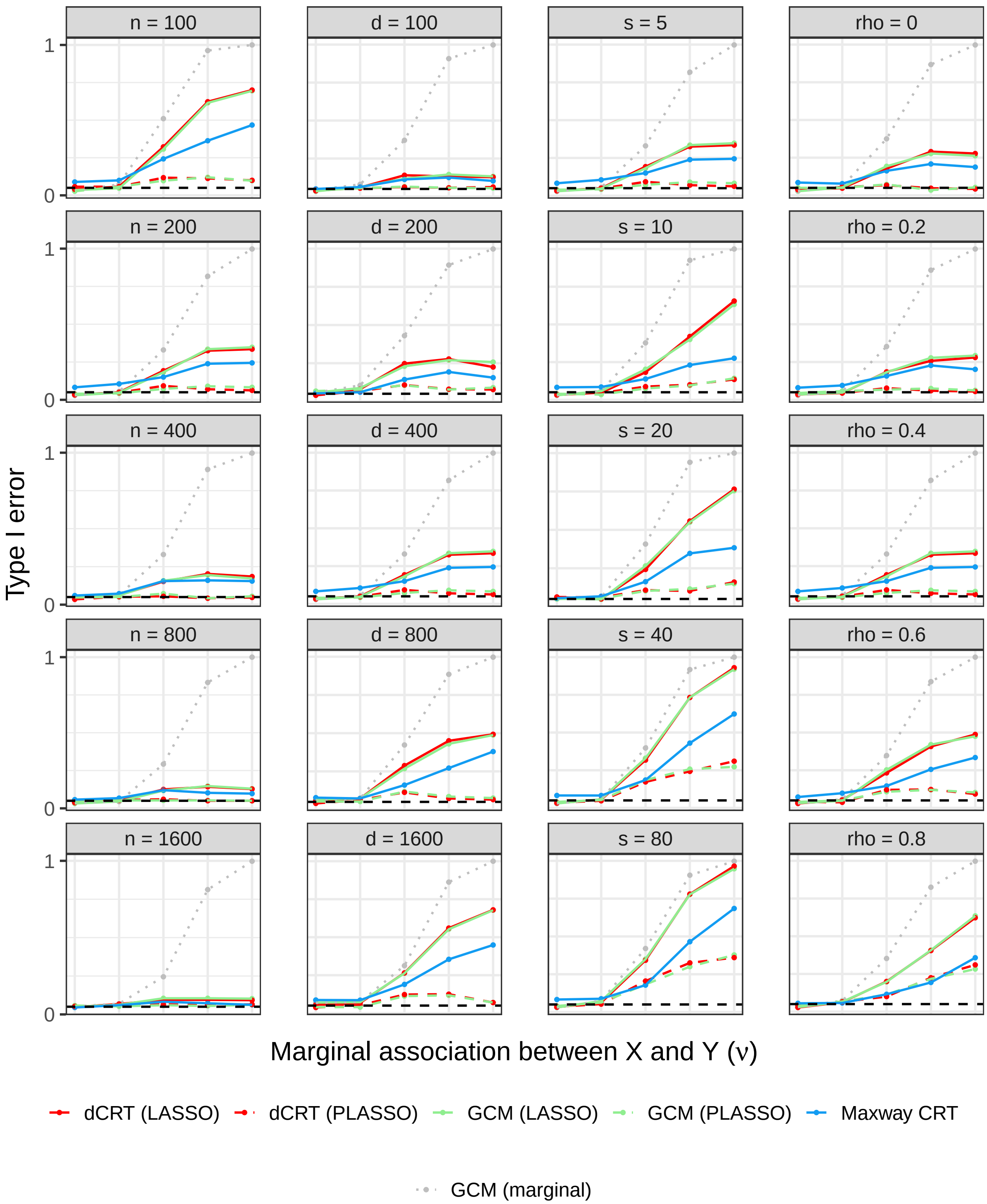

Figure 6: Type-I error in the Gaussian supervised setting.

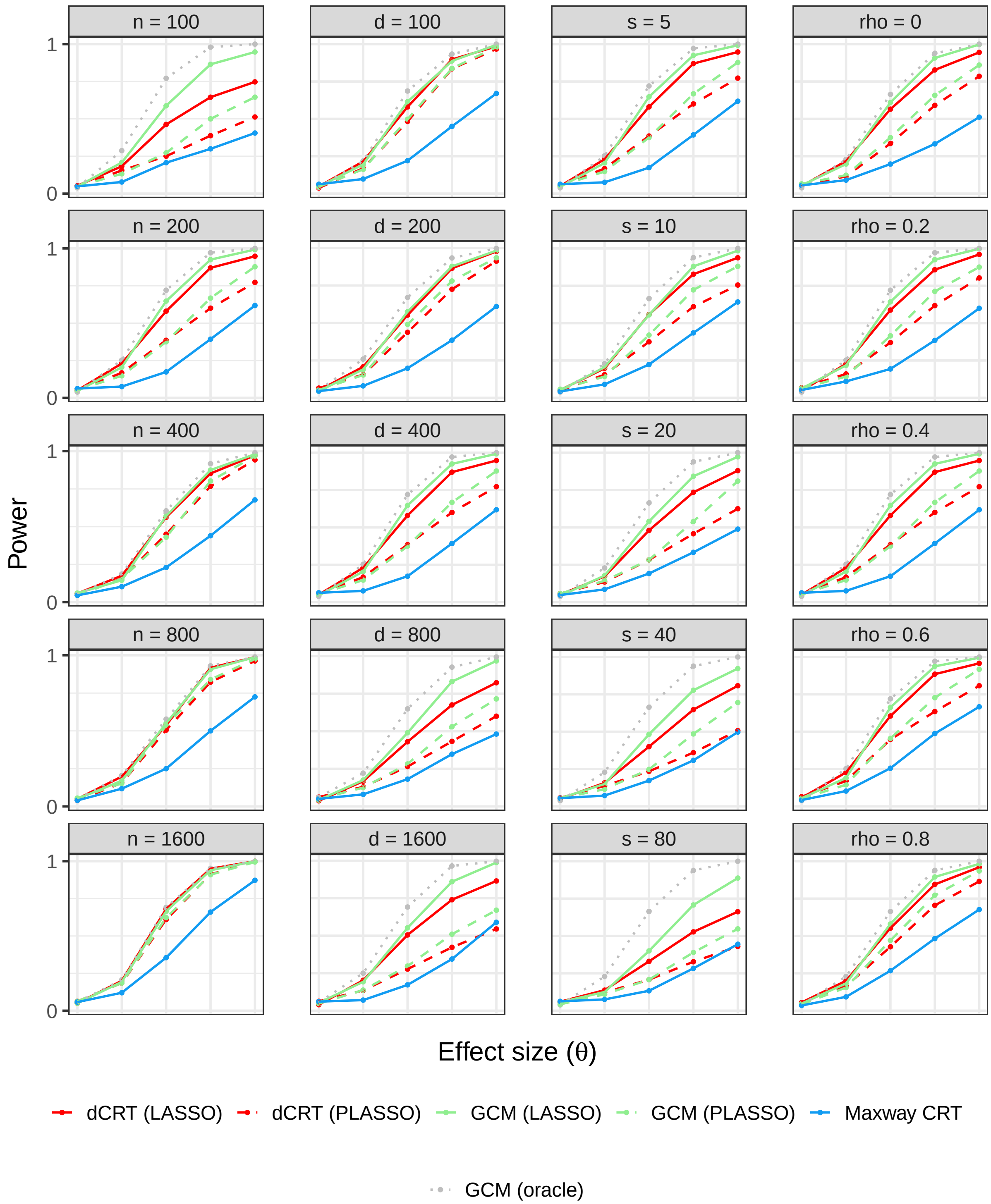

Figure 7: Power in the Gaussian supervised setting.

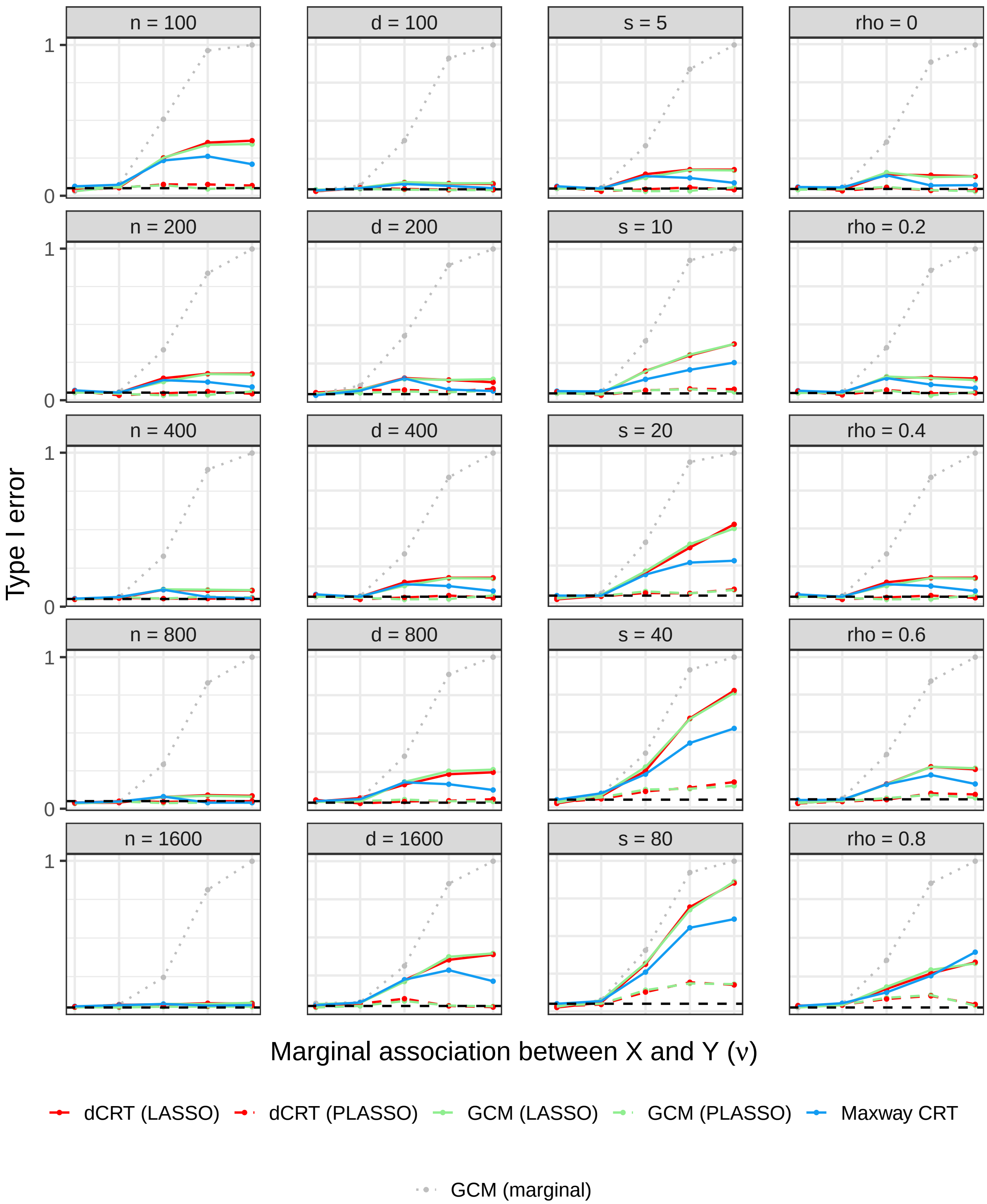

Figure 8: Type-I error in the Gaussian semi-supervised setting.

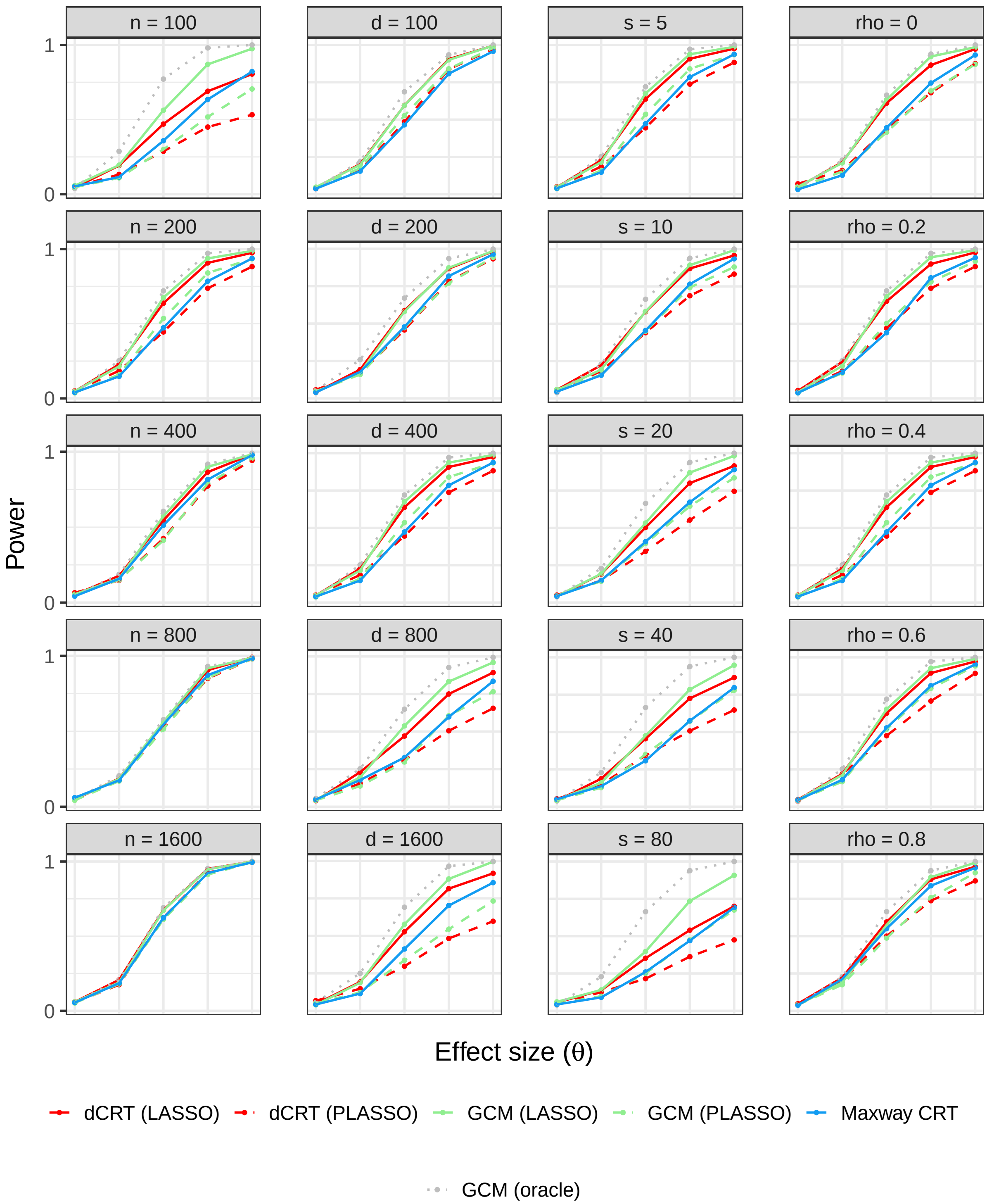

Figure 9: Power in the Gaussian semi-supervised setting.

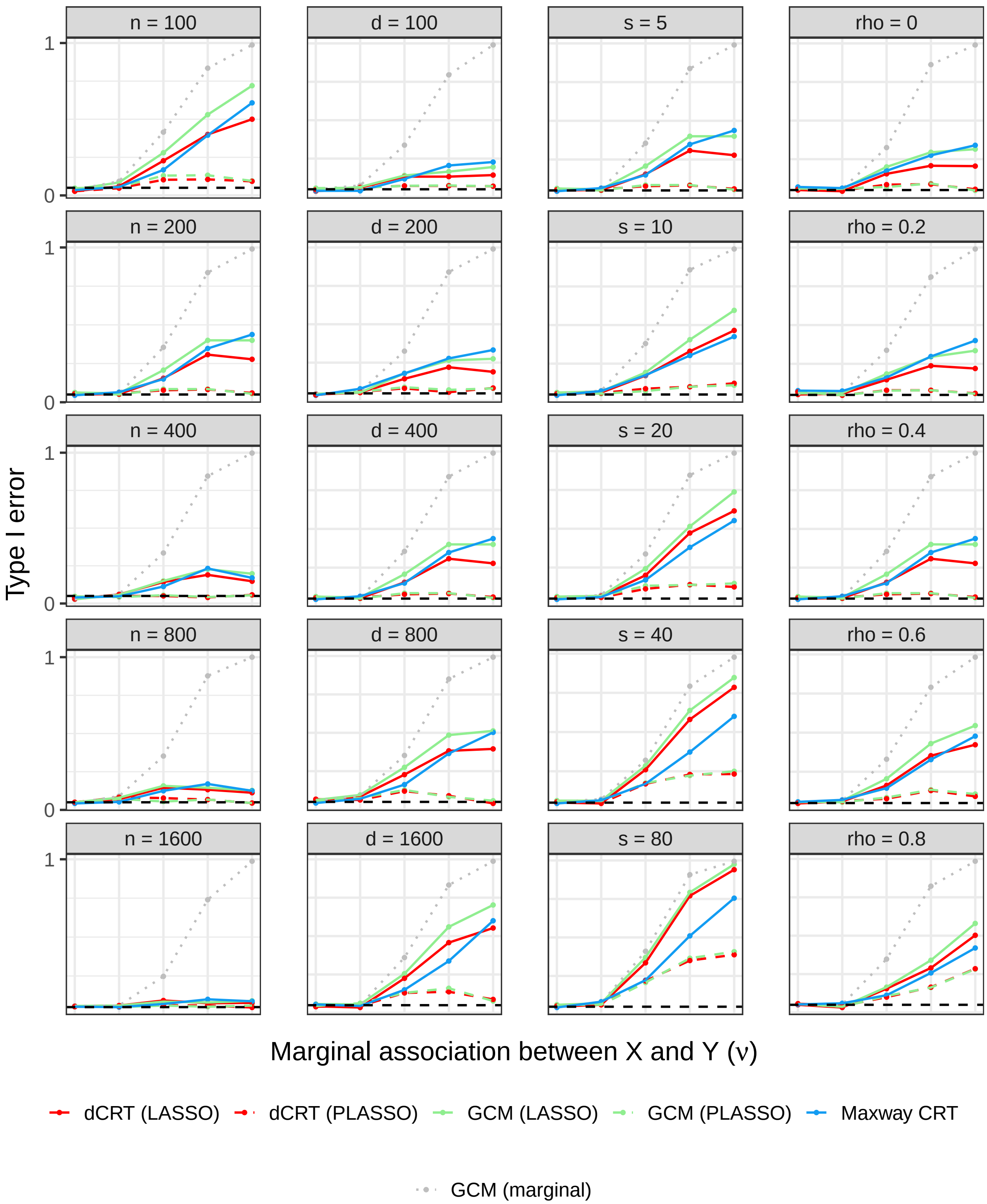

Figure 10: Type-I error in the binary supervised setting.

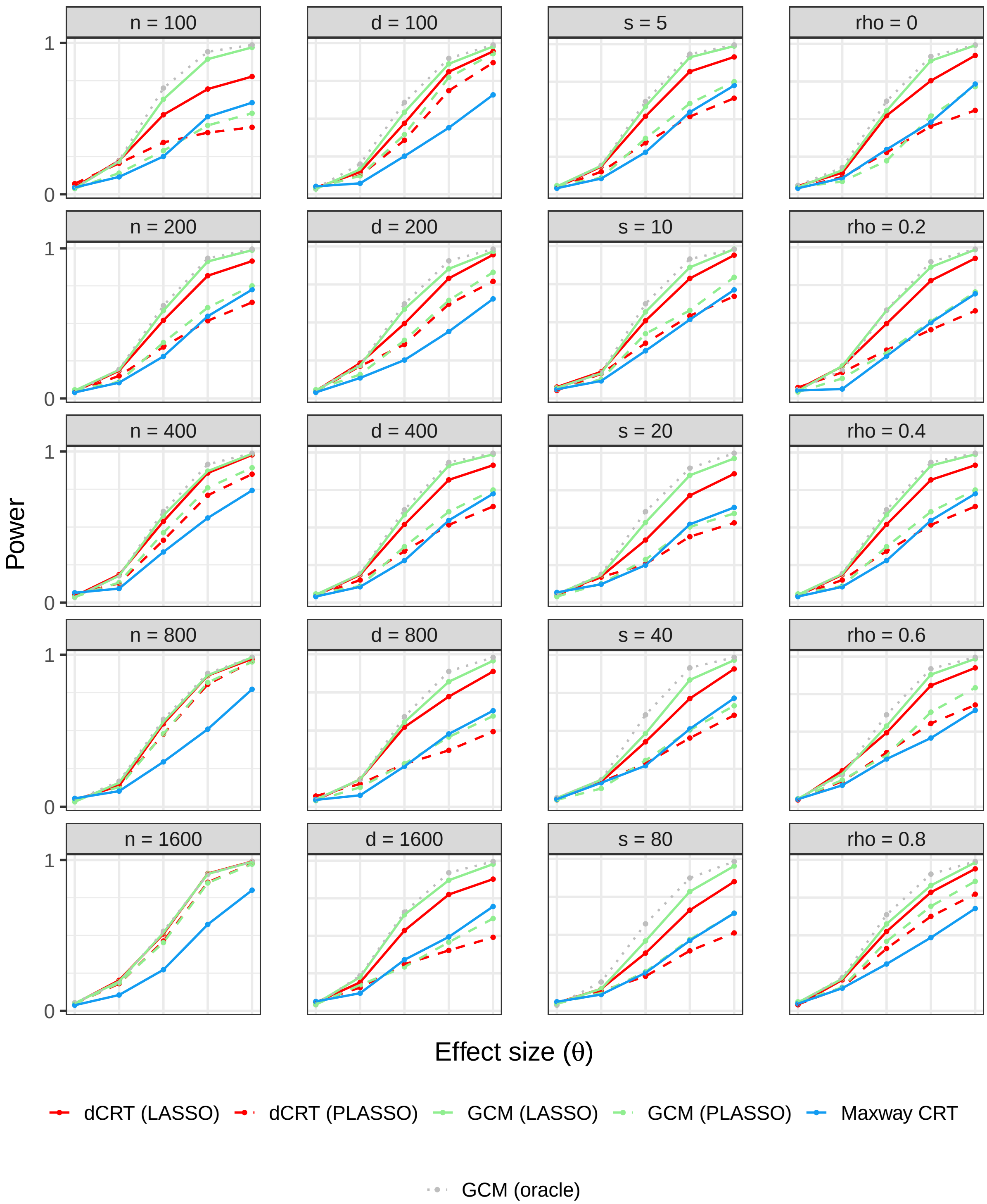

Figure 11: Power in the binary supervised setting.

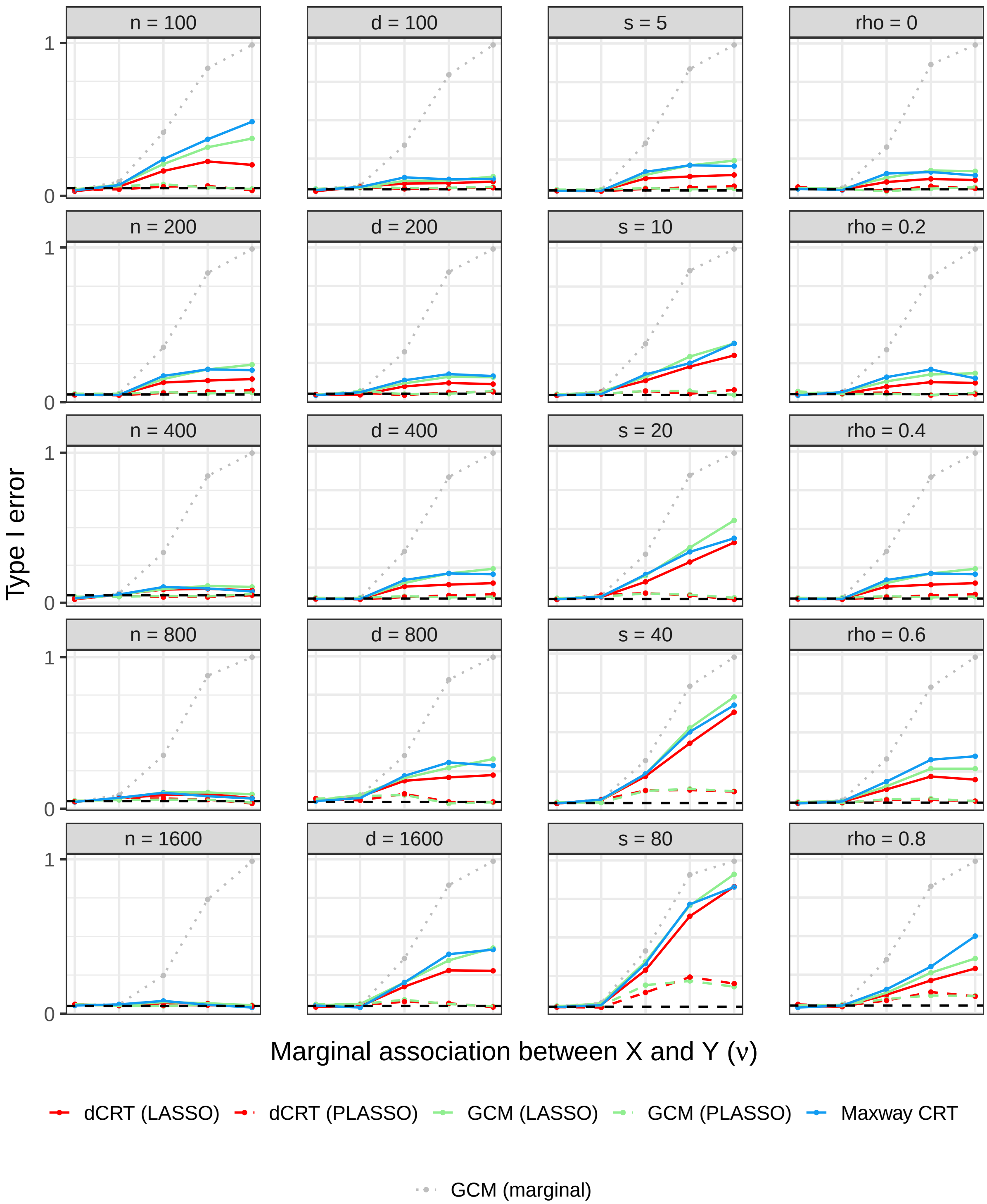

Figure 12: Type-I error in the binary semi-supervised setting.

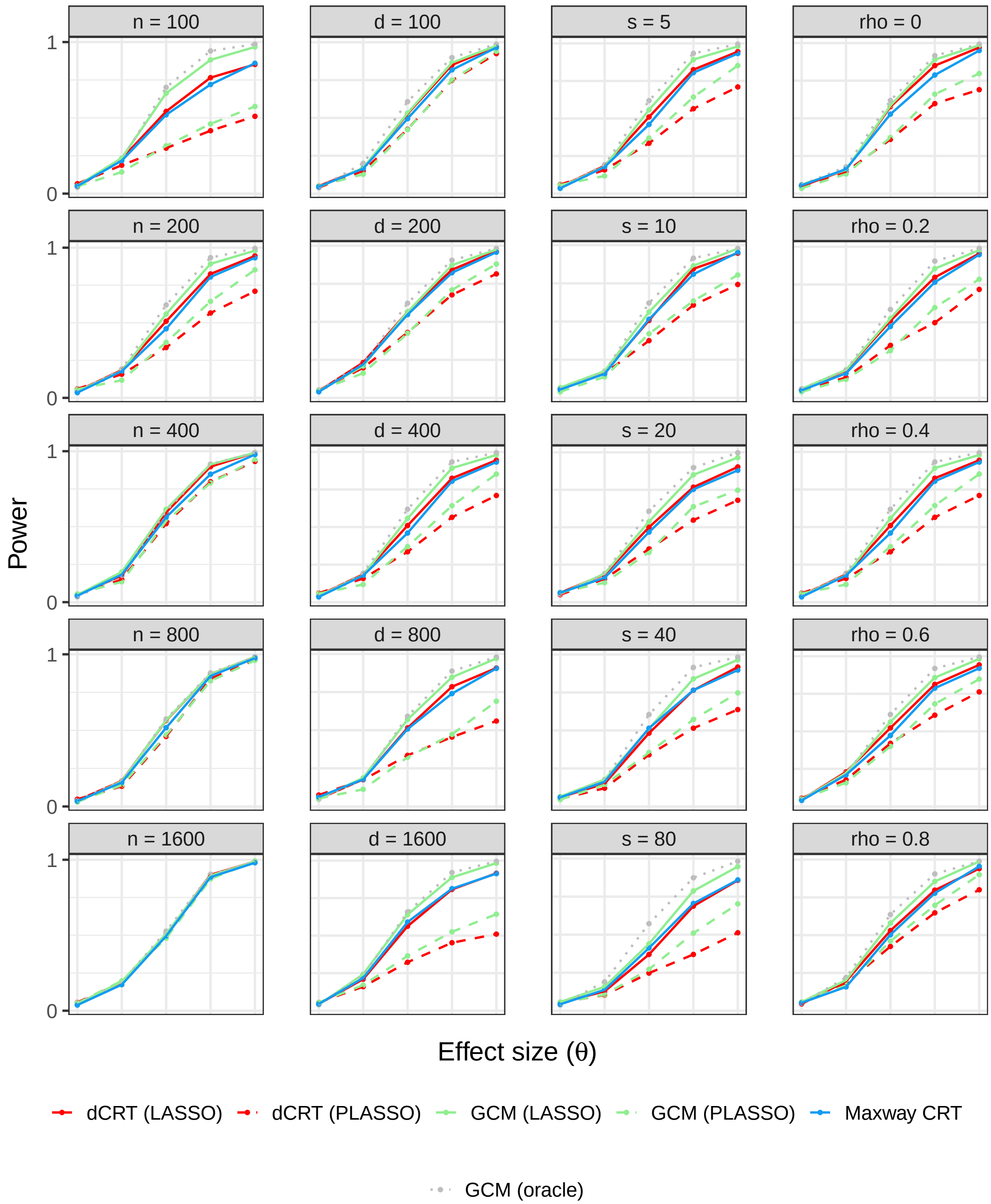

Figure 13: Power in the binary semi-supervised setting.

G Proofs of conditional convergence results

In this section, we present the proofs of the conditional convergence results from Appendix B. We proceed by stating and proving necessary lemmas in Section G.1 and then proving the convergence results themselves in Section G.2.

G.1 Auxiliary lemmas

First we state a few known results for the reader's convenience.

Lemma 10 (Durrett, 2010, Theorem 2.3.2). *A sequence of random variables W_n converges to a limit W in probability if and only if every subsequence of W_n has a further subsequence that converges to W almost surely.*

Lemma 11 (Conditional Markov inequality, Davidson, 2003, Theorem 10.17). *Let W be a random variable and let $\mathcal{F}$ be a σ-algebra. If for some $q > 0$ we have $\mathbb{E}[|W|^q] < \infty$, then for any ϵ we have*

$$\mathbb{P}(|W| \geq \epsilon | \mathcal{F}) \leq \frac{\mathbb{E}[|W|^q | \mathcal{F}]}{\epsilon^q} \quad \textit{almost surely.}$$

Lemma 12 (Conditional Hölder inequality, Swanson, 2019, Theorem 6.60). *Let W_1 and W_2 be random variables and let $\mathcal{F}$ be a σ-algebra. If for some $q_1, q_2 \in (1, \infty)$ with $\frac{1}{q_1} + \frac{1}{q_2} = 1$ we have $\mathbb{E}[|W_1|^{q_1}], \mathbb{E}[|W_2|^{q_2}] < \infty$, then*

$$\mathbb{E}[|W_1 W_2| \mid \mathcal{F}] \leq (\mathbb{E}[|W_1|^{q_1} \mid \mathcal{F}])^{1/q_1} (\mathbb{E}[|W_2|^{q_2} \mid \mathcal{F}])^{1/q_2} \quad \textit{almost surely.}$$

Lemma 13 (Romano and Lehmann, 2005, Lemma 11.2.1). *Suppose $W_n \xrightarrow{d} W$. If for given $\alpha \in (0, 1)$ the CDF of W is continuous and strictly increasing at $\mathbb{Q}_\alpha[W]$, then*

$$\mathbb{Q}_\alpha[W_n] \to \mathbb{Q}_\alpha[W].$$

Next we establish that, without of loss of generality, all random variables, σ-algebras, and conditional expectations in a triangular array may be viewed as being defined on a common probability space.

Lemma 14 (Embedding into a single probability space). *Consider a sequence of probability spaces $\{(\mathbb{P}_n, \Omega_n, \mathcal{G}_n), n \geq 1\}$. For each n, let $\{W_{i,n}\}_{i\geq 1}$ be a collection of integrable random variables defined on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ and let $\mathcal{F}_n \subseteq \mathcal{G}_n$ be a σ-algebra. Then there exists a single probability space $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$, random variables $\{\widetilde{W}_{i,n}\}_{i,n\geq 1}$ on $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$, and σ-fields $\widetilde{\mathcal{F}}_n \subseteq \widetilde{\mathcal{G}}$ for $n \geq 1$, such that for each n, the joint distribution of $(\{W_{i,n}\}_{i\geq 1}, \{\mathbb{E}[W_{i,n} \mid \mathcal{F}_n]\}_{i\geq 1})$ on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ coincides with that of $(\{\widetilde{W}_{i,n}\}_{i\geq 1}, \{\mathbb{E}[\widetilde{W}_{i,n} \mid \widetilde{\mathcal{F}}_n]\}_{i\geq 1})$ on $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$.*

Proof. Define the Cartesian product $\widetilde{\Omega} \equiv \prod_{n=1}^{\infty} \Omega_n$, the σ-algebra $\widetilde{\mathcal{G}}$ generated by measurable cylinders $\prod_{n=1}^{\infty} A_n$ for $A_n \in \mathcal{G}_n$ and $A_n = \mathcal{G}_n$ for all but finitely many n, and the infinite product measure $\widetilde{\mathbb{P}}$ on the measurable space $(\widetilde{\Omega}, \widetilde{\mathcal{G}})$ (Saeki, 1996). On this probability space, define σ-algebras

$$\widetilde{\mathcal{F}}_n \equiv \{\mathcal{G}_1 \times \cdots \times \mathcal{G}_{n-1} \times A_n \times \mathcal{G}_{n+1} \times \cdots : A_n \in \mathcal{F}_n\} \tag{134}$$

and random variables

$$\widetilde{W}_{in}(\omega) \equiv W_{in}(\omega_n) \tag{135}$$

for each $i, n \geq 1$. Next we claim that for each $i, n \geq 1$, the random variable

$$\mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n](\omega) \equiv \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n) \quad \text{for each } \omega \in \widetilde{\Omega} \tag{136}$$

is in fact a version of the conditional expectation $\mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n]$. Indeed, it suffices to check that for each $A \equiv \mathcal{G}_1 \times \cdots \times \mathcal{G}_{n-1} \times A_n \times \mathcal{G}_{n+1} \times \cdots \in \widetilde{\mathcal{F}}_n$ we have

$$\begin{aligned}
\int_A \mathbb{E}_{\widetilde{\mathbb{P}}}[\widetilde{W}_{in}|\widetilde{\mathcal{F}}_n](\omega) d\widetilde{\mathbb{P}}(\omega) &\equiv \int_A \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n) \mathrm{d}\widetilde{\mathbb{P}}(\omega) \\
&= \int_{\prod_{n' \neq n} \Omega_{n'}} \int_{A_n} \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n](\omega_n) \mathrm{d}\widetilde{\mathbb{P}}_n(\omega_n) \mathrm{d} \prod_{n' \neq n} \mathbb{P}_{n'}(\omega_{n'}) \\
&\equiv \int_{\prod_{n' \neq n} \Omega_{n'}} \int_{A_n} W_{in}(\omega_n) \mathrm{d}\widetilde{\mathbb{P}}_n(\omega_n) \mathrm{d} \prod_{n' \neq n} \mathbb{P}_{n'}(\omega_{n'}) \qquad (137) \\
&= \int_A W_{in}(\omega_n) \mathrm{d}\widetilde{\mathbb{P}}(\omega) \\
&\equiv \int_A \widetilde{W}_{in}(\omega) d\widetilde{\mathbb{P}}(\omega).
\end{aligned}$$

From the ω-wise embeddings (135) and (136), it is easy to verify the claimed equality between the joint distributions on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ and $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$. □

Finally, we state a conditional version of the truncated weak law of large numbers:

Lemma 15. *For each n, let $W_{in}, 1 \leq i \leq n$ be a set of random variables independent conditionally on $\mathcal{F}_n$. Let $b_n > 0$ with $b_n \to \infty$ and let $\bar{W}_{in} = W_{in}\mathbb{1}(|W_{in}| \leq b_n)$. Suppose that as $n \to \infty$ we have*

1. $\sum_{i=1}^n \mathbb{P}[|W_{in}| > b_n|\mathcal{F}_n] \xrightarrow{p} 0$ *and*
2. $b_n^{-2} \sum_{i=1}^n \mathbb{E}[\bar{W}_{in}^2|\mathcal{F}_n] \xrightarrow{p} 0.$

If we set $S_n \equiv \sum_{i=1}^n W_{in}$ and $a_n \equiv \sum_{i=1}^n \mathbb{E}[\bar{W}_{in}]$ then

$$\frac{S_n - a_n}{b_n} \mid \mathcal{F}_n \xrightarrow{p,p} 0.$$

Proof. Let $\bar{S}_n \equiv \sum_{i=1}^n \bar{W}_{in}$. We first write

$$\mathbb{P}\left[\left|\frac{S_n - a_n}{b_n}\right| > \epsilon \middle| \mathcal{F}_n\right] \leq \mathbb{P}\left[S_n \neq \bar{S}_n | \mathcal{F}_n\right] + \mathbb{P}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon \middle| \mathcal{F}_n\right].$$

To estimate the first term, we note that

$$\mathbb{P}[S_n \neq \bar{S}_n | \mathcal{F}_n] \leq \mathbb{P}[\cup_{i=1}^n \{\bar{W}_{in} \neq W_{in}\} | \mathcal{F}_n] \leq \sum_{i=1}^n \mathbb{P}(|W_{in}| > b_n | \mathcal{F}_n) \xrightarrow{p} 0$$

by the first assumption. For the second term, we note that conditional Markov's inequality (Lemma 11), $a_n = \mathbb{E}[\bar{S}_n | \mathcal{F}_n]$ implies that

$$\begin{aligned}
\mathbb{P}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right| > \epsilon | \mathcal{F}_n\right] &\leq \epsilon^{-2} \mathbb{E}\left[\left|\frac{\bar{S}_n - a_n}{b_n}\right|^2 \middle| \mathcal{F}_n\right] \\
&= \epsilon^{-2} b_n^{-2} \mathrm{Var}[\bar{S}_n | \mathcal{F}_n] \\
&= (b_n \epsilon)^{-2} \sum_{i=1}^n \mathrm{Var}[\bar{W}_{in} | \mathcal{F}_n] \\
&\leq (b_n \epsilon)^{-2} \sum_{i=1}^n \mathbb{E}\left[\bar{X}_{in}^2 | \mathcal{F}_n\right] \xrightarrow{p} 0,
\end{aligned}$$

where the convergence in the last line is given by the second assumption. This completes the proof. □

G.2 Proofs of conditional convergence results

Proof of Theorem 5. This proof generalizes the argument of Van Der Vaart, 1998, Lemma 2.11 to allow for conditioning. Fix $\epsilon > 0$, and choose an integer $k \geq 2/\epsilon$. Because the CDF of W is continuous, it follows that

$$\mathbb{P}[W \leq \mathbb{Q}_{i/k}[W]] = i/k \quad \text{for each } 0 \leq i \leq k.$$

Fix $t \in \mathbb{R}$, and suppose $\mathbb{Q}_{\frac{i-1}{k}}[W] \leq t \leq \mathbb{Q}_{\frac{i}{k}}[W]$. It follows that

$$\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i-1}{k}}[W] \mid \mathcal{F}_n] - \frac{i}{k} \leq \mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t] \leq \mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \frac{i-1}{k}.$$

Therefore, for all $t \in \mathbb{R}$, we have

$$\begin{aligned}
|\mathbb{P}[W_n \leq t \mid \mathcal{F}_n] - \mathbb{P}[W \leq t]| &\leq \sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{i/k}[W] \mid \mathcal{F}_n] - \frac{i}{k}\right| + \frac{1}{k} \\
&= \sup_{0 \leq i \leq k} \left|\mathbb{P}[W_n \leq \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \leq \mathbb{Q}_{\frac{i}{k}}[W]]\right| + \frac{1}{k},
\end{aligned}$$

so that

$$\sup_{t\in\mathbb{R}} |\mathbb{P}[W_n \le t \mid \mathcal{F}_n] - \mathbb{P}[W \le t]| \le \sup_{0\le i\le k} \left|\mathbb{P}[W_n \le \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \le \mathbb{Q}_{\frac{i}{k}}[W]]\right| + \frac{1}{k}.$$

By assumption, we have

$$\mathbb{P}\left[\left|\mathbb{P}[W_n \le \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \le \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k(k+1)}\right] \to 0. \tag{138}$$

Therefore,

$$\begin{aligned}
&\mathbb{P}\left[\sup_{t\in\mathbb{R}} |\mathbb{P}[W_n \le t \mid \mathcal{F}_n] - \mathbb{P}[W \le t]| > \epsilon\right] \\
&\le \mathbb{P}\left[\sup_{t\in\mathbb{R}} |\mathbb{P}[W_n \le t \mid \mathcal{F}_n] - \mathbb{P}[W \le t]| > \frac{2}{k}\right] \\
&\le \mathbb{P}\left[\sup_{0\le i\le k} \left|\mathbb{P}[W_n \le \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \le \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k}\right] \\
&\le \sum_{i=0}^{k} \mathbb{P}\left[\left|\mathbb{P}[W_n \le \mathbb{Q}_{\frac{i}{k}}[W] \mid \mathcal{F}_n] - \mathbb{P}[W \le \mathbb{Q}_{\frac{i}{k}}[W]]\right| > \frac{1}{k(k+1)}\right] \\
&\to 0.
\end{aligned}$$

This completes the proof. □

Proof of Theorem 6. Fix $t \in \mathbb{R}$. Letting $F(t') \equiv \mathbb{P}[W \le t']$ be the CDF of W, Theorem 5 gives

$$\left|\mathbb{P}\left[W_n \le \frac{t-b_n}{a_n} \mid \mathcal{F}_n\right] - F\left(\frac{t-b_n}{a_n}\right)\right| \le \sup_{t'\in\mathbb{R}} |\mathbb{P}[W_n \le t' | \mathcal{F}_n] - F(t')| \xrightarrow{p} 0.$$

By the continuous mapping theorem, we have $F\left(\frac{t-b_n}{a_n}\right) \xrightarrow{p} F(t)$, so that

$$\mathbb{P}\left[W_n \le \frac{t-b_n}{a_n} \mid \mathcal{F}_n\right] \xrightarrow{p} F(t).$$

Noting that $\mathbb{P}[a_n \le 0 | \mathcal{F}_n]$ is a sequence of nonnegative random variables whose expectations converge to zero, it follows that $\mathbb{P}[a_n \le 0|\mathcal{F}_n] \xrightarrow{p} 0$ and so

$$\begin{aligned}
\mathbb{P}\left[a_n W_n + b_n \le t \mid \mathcal{F}_n\right] &= \mathbb{P}\left[a_n W_n + b_n \le t, a_n > 0 \mid \mathcal{F}_n\right] + o_p(1) \\
&= \mathbb{P}\left[W_n \le \frac{t-b_n}{a_n}, a_n > 0 \mid \mathcal{F}_n\right] + o_p(1) \\
&= \mathbb{P}\left[W_n \le \frac{t-b_n}{a_n} \mid \mathcal{F}_n\right] + o_p(1) \\
&\xrightarrow{p} F(t),
\end{aligned}$$

as desired. □

Proof of Theorem 7. We apply Lemma 15 with $b_n = n$. We first verify the first assumption in Lemma 15 by conditional Markov's inequality (Lemma 11):

$$\sum_{i=1}^{n} \mathbb{P}(|W_{in}| > n|\mathcal{F}_n) \le \sum_{i=1}^{n} \frac{\mathbb{E}(|W_{in}|^{1+\delta}|\mathcal{F}_n)}{n^{1+\delta}} \xrightarrow{p} 0.$$

For the second condition, we have

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}[W_{in}^2 \mathbb{1}(|W_{in}| \le n)|\mathcal{F}_n] &\le \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} n^{1-\delta} \mathbb{1}(|W_{in}| \le n)] \\
&= \frac{1}{n^{1+\delta}}\sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} \mathbb{1}(|W_{in}| \le n)|\mathcal{F}_n] \\
&\le \frac{1}{n^{1+\delta}}\sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n] \xrightarrow{p} 0.
\end{aligned}$$

Therefore, Lemma 15 yields

$$\frac{1}{n}\sum_{i=1}^{n}(W_{in} - \mathbb{E}[W_{in}\mathbb{1}(|W_{in}| \le n) \mid \mathcal{F}_n]) \;\Bigg|\; \mathcal{F}_n \xrightarrow{p,p} 0.$$

By conditional Slutsky (Theorem 6), it now suffices to show that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[W_{in}\mathbb{1}(|W_{in}| > n) \mid \mathcal{F}_n\right] \xrightarrow{p} 0.$$

To see this, applying conditional Markov's and Hölder's inequalities (Lemmas 11 and 12, respectively) we obtain

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[W_{in}\mathbb{1}(|W_{in}| > n)|\mathcal{F}_n\right]\right| &\le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[|W_{in}|\mathbb{1}(|W_{in}| > n)|\mathcal{F}_n\right] \\
&\le \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n]\}^{1/(1+\delta)}\{\mathbb{P}[|W_{in}| > n|\mathcal{F}_n]\}^{\delta/(1+\delta)} \\
&\le \frac{1}{n}\sum_{i=1}^{n}\{\mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n]\}^{1/(1+\delta)}\left\{\frac{\mathbb{E}(|W_{in}|^{1+\delta}|\mathcal{F}_n)}{n^{1+\delta}}\right\}^{\delta/(1+\delta)} \\
&= \frac{1}{n^{1+\delta}}\sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta}|\mathcal{F}_n] \\
&\xrightarrow{p} 0,
\end{aligned}$$

where the last convergence is by assumption. Finally, we verify that the condition (57) is sufficient for the conditional WLLN assumption (55) by noting that it implies

$$\frac{1}{n^{1+\delta}}\sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n] \le \frac{\sup_{1\le i\le n}\mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n]n}{n^{1+\delta}} = \frac{\sup_{1\le i\le n}\mathbb{E}[|W_{in}|^{1+\delta} \mid \mathcal{F}_n]}{n^{\delta}} \xrightarrow{p} 0.$$

This completes the proof. $\square$

Proof of Theorem 8. Without loss of generality, we assume $\mathbb{E}[W_{in}|\mathcal{F}_n] = 0$ and that all random variables and σ-algebras are defined on a common probability space $(\mathbb{P}, \Omega, \mathcal{G})$ (Lemma 14). Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra on $\mathbb{R}^n$. Let κ_n be a regular conditional distribution of $(W_{1n}, \ldots, W_{nn})$ given $\mathcal{F}_n$ (Klenke, 2017, Theorem 8.37), i.e. a function $\kappa_n : \Omega \times \mathcal{B}(\mathbb{R}^n) \to [0, \infty]$ such that $\omega \mapsto \kappa_n(\omega, B)$ is measurable for each $B \in \mathcal{B}(\mathbb{R}^n)$, $B \mapsto \kappa_n(\omega, B)$ is a σ-finite measure on $\mathbb{R}^n$ for each $\omega \in \Omega$, and

$$\kappa_n(\omega, B) = \mathbb{P}[(W_{1n}, \ldots, W_{nn}) \in B|\mathcal{F}_n](\omega), \quad \text{for almost all } \omega \in \Omega \text{ and all } B \in \mathcal{B}(\mathbb{R}^n).$$

For each n and each $\omega \in \Omega$, let $(\widetilde{W}_{1n}(\omega), \ldots, \widetilde{W}_{nn}(\omega))$ be a draw from the measure $\kappa_n(\omega, \cdot)$. By Klenke (2017, Theorem 8.38), we have for each n that

$$\left(\sum_{i=1}^{n} \mathrm{Var}[\widetilde{W}_{in}(\omega)]\right)^{-(2+\delta)/2} \sum_{i=1}^{n} \mathbb{E}[|\widetilde{W}_{in}(\omega)|^{2+\delta}] \overset{a.s.}{=} \frac{1}{S_n^{2+\delta}(\omega)} \sum_{i=1}^{n} \mathbb{E}[|W_{in}|^{2+\delta} \mid \mathcal{F}_n](\omega).$$

Now, let $\{n_k\}_{k\geq 1}$ be a subsequence of $\mathbb{N}$. By the conditional Lyapunov assumption (62) and Lemma 10, there is a further subsequence n_{k_j} such that

$$\frac{1}{S_{n_{k_j}}^{2+\delta}} \sum_{i=1}^{n_{k_j}} \mathbb{E}[|W_{in_{k_j}}|^{2+\delta} \mid \mathcal{F}_{n_{k_j}}] \overset{a.s.}{\to} 0. \tag{139}$$

Hence, it follows that

$$\left(\sum_{i=1}^{n_{k_j}} \mathrm{Var}[\widetilde{W}_{in_{k_j}}(\omega)]\right)^{-(2+\delta)/2} \sum_{i=1}^{n_{k_j}} \mathbb{E}[|\widetilde{W}_{in_{k_j}}(\omega)|^{2+\delta}] \to 0 \quad \text{for almost every } \omega \in \Omega. \tag{140}$$

Applying the usual Lyapunov CLT to the triangular array $\{\widetilde{W}_{in_{k_j}}(\omega)\}_{i,n_{k_j}}$, we find that

$$\left(\sum_{i=1}^{n_{k_j}} \mathrm{Var}[\widetilde{W}_{in_{k_j}}(\omega)]\right)^{-1} \sum_{i=1}^{n_{k_j}} \widetilde{W}_{in_{k_j}}(\omega) \overset{d}{\to} N(0,1) \quad \text{for almost every } \omega \in \Omega, \tag{141}$$

and therefore that, for each $t \in \mathbb{R}$, we have

$$\mathbb{P}\left[\left(\sum_{i=1}^{n_{k_j}} \mathrm{Var}[\widetilde{W}_{in_{k_j}}(\omega)]\right)^{-1} \sum_{i=1}^{n_{k_j}} \widetilde{W}_{in_{k_j}}(\omega) \leq t\right] \to \Phi(t) \quad \text{for almost every } \omega \in \Omega. \tag{142}$$

Using Klenke (2017, Theorem 8.38) again, it follows that for each $t \in \mathbb{R}$, we have

$$\mathbb{P}\left[\frac{1}{S_{n_{k_j}}} \sum_{i=1}^{n_{k_j}} W_{in_{k_j}} \leq t \mid \mathcal{F}_{n_{k_j}}\right] \overset{a.s.}{\to} \Phi(t). \tag{143}$$

Applying Lemma 10, it follows that

$$\mathbb{P}\left[\frac{1}{S_n}\sum_{i=1}^{n} W_{in} \leq t \mid \mathcal{F}_n\right] \xrightarrow{p} \Phi(t), \tag{144}$$

as desired. $\square$

Proof of Lemma 1. Without loss of generality, we assume that all random variables and σ-algebras are defined on a common probability space $(\mathbb{P}, \Omega, \mathcal{G})$ (Lemma 14). Let $\mathcal{B}(\mathbb{R})$ be the Borel σ-algebra on $\mathbb{R}$. Let κ_n be a regular conditional distribution of W_n given $\mathcal{F}_n$ (Klenke, 2017, Theorem 8.29), i.e. a function $\kappa_n : \Omega \times \mathcal{B}(\mathbb{R}) \to [0, \infty]$ such that $\omega \mapsto \kappa_n(\omega, B)$ is measurable for each $B \in \mathcal{B}(\mathbb{R})$, $B \mapsto \kappa_n(\omega, B)$ is a σ-finite measure on $\mathbb{R}^n$ for each $\omega \in \Omega$, and

$$\kappa_n(\omega, B) = \mathbb{P}[W_n \in B|\mathcal{F}_n](\omega), \quad \text{for almost all } \omega \in \Omega \text{ and all } B \in \mathcal{B}(\mathbb{R}).$$

Now, let $\{n_k\}_{k\geq 1}$ be a subsequence of $\mathbb{N}$. By conditional Polya's theorem (Theorem 5), we have

$$\sup_{t\in\mathbb{R}} |\mathbb{P}[W_n \leq t|\mathcal{F}_n] - \mathbb{P}[W \leq t]| \xrightarrow{p} 0. \tag{145}$$

Hence, by Lemma 10 there is a further subsequence n_{k_j} such that

$$\sup_{t\in\mathbb{R}} |\mathbb{P}[W_{n_{k_j}} \leq t|\mathcal{F}_{n_{k_j}}](\omega) - \mathbb{P}(W \leq t)| \to 0, \text{ for almost all } \omega \in \Omega. \tag{146}$$

It follows that

$$\sup_{t\in\mathbb{R}} |\kappa_{n_{k_j}}(\omega, (-\infty, t]) - \mathbb{P}(W \leq t)| \to 0, \text{ for almost all } \omega \in \Omega, \tag{147}$$

i.e.

$$\kappa_{n_{k_j}}(\omega, \cdot) \xrightarrow{d} W \text{ for almost all } \omega \in \Omega. \tag{148}$$

Hence, by Lemma 13, it follows that

$$\mathbb{Q}_\alpha[\kappa_{n_{k_j}}(\omega, \cdot)] \to \mathbb{Q}_\alpha[W] \text{ for almost all } \omega \in \Omega, \tag{149}$$

and therefore

$$\mathbb{Q}_\alpha[W_{n_{k_j}}|\mathcal{F}_{n_{k_j}}] \xrightarrow{a.s.} \mathbb{Q}_\alpha[W]. \tag{150}$$

Applying Lemma 10 again, we conclude that

$$\mathbb{Q}_\alpha[W_n|\mathcal{F}_n] \xrightarrow{p} \mathbb{Q}_\alpha[W], \tag{151}$$

as desired. $\square$