The conditional saddlepoint approximation for fast and accurate large-scale hypothesis testing

Ziang Niu, Jyotishka Ray Choudhury, Eugene Katsevich

March 30, 2025

Abstract

Conditional independence (CI) testing is a fundamental problem in statistics. Modern data often exhibit large scale and high sparsity. CI testing methods using asymptotic normal approximations can induce inflated Type-I error rates when dealing with sparse data. Resampling-based CI testing procedures, while statistically more accurate, are computationally expensive, making them impractical for large-scale problems. To resolve this dilemma, we rely on the saddlepoint approximation (SPA) technique and establish a new SPA for the conditional distribution of a resampled test statistic that can be expressed as a sample average. The new result generalizes the existing SPA results in multiple facets and provide the first satisfactory justification of SPA for sign-flipping testing procedure proposed 70 years ago (Daniels, 1955). Leveraging the new theoretical result, we introduce the saddlepoint approximation-based conditional randomization test (spaCRT), a novel conditional independence test that effectively balances statistical accuracy and computational efficiency. The validity of spaCRT is assessed with modern regression techniques applied such as lasso and kernel ridge regression. Through extensive simulations and real data analysis, we demonstrate that the spaCRT controls Type-I error and maintains high power, outperforming other asymptotic and resampling-based tests. Our method is particularly well-suited for data with extreme sparsity and large scale, such as single-cell CRISPR screen analyses and GWASs involving rare diseases.

1 Introduction

Given the joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}_n$ (potentially depending on sample size n), the statistical task of conditional independence testing is to test the null hypothesis

$$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}. \tag{1}$$

In the language of causal inference (Imbens and Rubin, 2015), $\boldsymbol{X}$ is often interpreted as a treatment variable, $\boldsymbol{Y}$ as an outcome variable, and $\boldsymbol{Z}$ as a set of covariates. The null hypothesis H_0 states that the treatment has no effect on the outcome, given the covariates. Suppose the independent and identically distributed observations $(X_{in}, Y_{in}, Z_{in}) \overset{\text{i.i.d.}}{\sim} \mathcal{L}_n$ are collected for subject $i = 1, \ldots, n$. We denote these observations collectively as $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times d}$.

1.1 Challenges: data sparsity and large scale

A particular challenge in modern applications involving tests of the CI null (1) is the high sparsity of the data—that is, an excessive number of zeros in the observations. This issue is especially prevalent in genetic and genomic data analysis. Moreover, sparsity is often accompanied by large scale, which further hinders the use of sophisticated but computationally intensive statistical methods.

A concrete example arises in population genetics. Genome-wide association studies (GWASs) investigate the relationship between genetic variants ($\boldsymbol{X}$) and human diseases ($\boldsymbol{Y}$) (Visscher et al., 2017), while accounting for potential confounding variables ($\boldsymbol{Z}$), such as population structure and experimental covariates like age and sex. Whole-genome sequencing can generate genotype data across thousands or even tens of thousands of loci for large population cohorts. When studying rare diseases and rare genetic variants, the observed data X and Y often contain an excessive number of zeros. These two sources of sparsity have been widely recognized as major challenges in GWASs (Auer and Lettre, 2015; Dey et al., 2017; Zhao et al., 2020; Turro et al., 2020).

Another example comes from single-cell CRISPR screens (Dixit et al., 2016; Adamson et al., 2016; Jaitin et al., 2016; Datlinger et al., 2017). One of the primary biological objectives in these experiments is to use a large number of *CRISPR perturbations* ($\boldsymbol{X}$) and gene expression measurements ($\boldsymbol{Y}$) at the single-cell level to determine, for each perturbation–gene pair, whether cells carrying the perturbation exhibit different expression levels compared to cells without the perturbation. The analysis is influenced by technical covariates ($\boldsymbol{Z}$), such as library size and batch effects. The number of hypotheses can reach hundreds of thousands of perturbation–gene pairs, with over 200,000 cells in a single experiment (Gasperini et al., 2019). Excessive sparsity in the treatment indicators X is a common challenge due to the pooled nature of large-scale perturbation libraries. In addition, the single-cell resolution of gene expression data often leads to sparse, count-based observations in Y.

Applying classical asymptotic inference methods in these sparse settings is often problematic, despite their computational appeal. Empirical evidence from both GWASs (Dey et al., 2017) and single-cell CRISPR screens (Barry et al., 2024) has shown inflated Type-I error rates when using tests based on the central limit theorem. A natural alternative is to employ more well-calibrated, resampling-based testing procedures. However, the large scale—both in sample size and in the number of hypotheses—makes the computational cost of such methods prohibitively high. As a result, direct applications of resampling techniques, such as the bootstrap or permutation tests, are often infeasible. This presents a dilemma for practitioners: how can one achieve accurate statistical inference while maintaining computational tractability in large-scale, high-sparsity scenarios?

1.2 Relevant literature

The tension between statistical accuracy and computational efficiency is not unique to CI testing; it is a common theme across many statistical applications. A substantial body of literature has been developed to address this challenge. We categorize the relevant works into two groups: resampling-based procedures and resampling-free procedures.

Resampling-based procedures have been accelerated through adaptive resampling schemes, which dynamically adjust the number of resamples based on the data (Besag and Clifford, 1991; Gandy, 2009; Gandy and Hahn, 2014; Gandy and Hahn, 2016; Gandy and Hahn, 2017; Fischer and Ramdas, 2024b; Fischer and Ramdas, 2024a). These methods are broadly applicable to arbitrary resampling schemes and test statistics, though they typically require

some resampling overhead. On the more applied side, to mitigate the computational burden of resampling, Ge et al. (2012) and Winkler et al. (2016) proposed heuristic approaches that fit parametric curves to the resampled distributions, allowing a relatively small number of resamples to be used for learning the curve. Barry et al. (2021) proposed the use of the dCRT (Liu et al., 2022)—an accelerated variant of the resampling-based conditional randomization test (CRT; Candès et al., 2018)—in the context of single-cell CRISPR screens, and applied a similar parametric curve-fitting technique to the resampling distribution of dCRT.

On the other hand, a large class of methods addressing this dilemma relies on finer asymptotic expansion techniques. For instance, the classical Edgeworth expansion has been used to improve upon the naïve normal approximation by better capturing the null distribution of test statistics (Hall, 2013; Bentkus, Götze, and Zwet, 1997; Bickel, 1974). However, its accuracy can deteriorate in the tail regions of the distribution (Hall, 2013), which poses a challenge in settings where small p-values are critical. The saddlepoint approximation (SPA; Daniels, 1954; Lugannani and Rice, 1980) is another classical technique that provides highly accurate approximations of both densities and tail probabilities. A key advantage of SPA is its relative error guarantee (Butler, 2007; Kolassa, 2006), which is especially important in large-scale settings where small p-values are common due to high multiplicity. SPA has also been proposed as an approximation for the resampling distributions of classical resampling-based procedures, such as permutation tests (Robinson, 1982) and the bootstrap (Davison and Hinkley, 1988), although the theoretical justification for these applications remains incomplete.

Despite the extensive literature discussed above, the challenge of reconciling statistical accuracy with computational efficiency remains unresolved in CI testing.

1.3 Our contributions

To fill this gap in CI testing, we propose a new approach that reconciles statistical accuracy with computational efficiency by leveraging the saddlepoint approximation (SPA) for resampling-based procedures. However, SPA has not been rigorously justified for any resampling-based methods, nor has it been applied in the context of CI testing. To address these challenges, we make two central contributions:

1. We first establish theoretical justification for SPA in the context of resampling-based hypothesis testing, where conditioning must be accounted for. This justifies the SPA for classical resampling-based procedures, like the sign-flipping test, 70 years after these approximations were first proposed. It also loosens the assumptions of SPA, not requiring continuity or lattice assumptions, establishing a new result even for the classical (unconditional) SPA.

2. To extend this useful approximation to CI testing, we apply the SPA to the dCRT to arrive at the *saddlepoint approximation-based conditional randomization test* (spaCRT), the first SPA for a CI testing procedure. We provide theoretical justification for the spaCRT in general, and in a variety of specific modern settings encompassing high-dimensional and nonparametric machine learning regressions. We provide the R package spacrt to implement the spaCRT, which is available on GitHub.

Through simulation studies inspired by single-cell CRISPR screen and GWAS applications, we demonstrate that spaCRT provides fast and accurate inference. We further apply spaCRT to a real single-cell CRISPR screen dataset, achieving an approximately 250-fold

speedup. The results, previewed in Figure 1, highlight that spaCRT offers substantially improved Type I error control compared to the normal approximation-based GCM test, while being significantly faster than dCRT and nearly as fast as GCM.

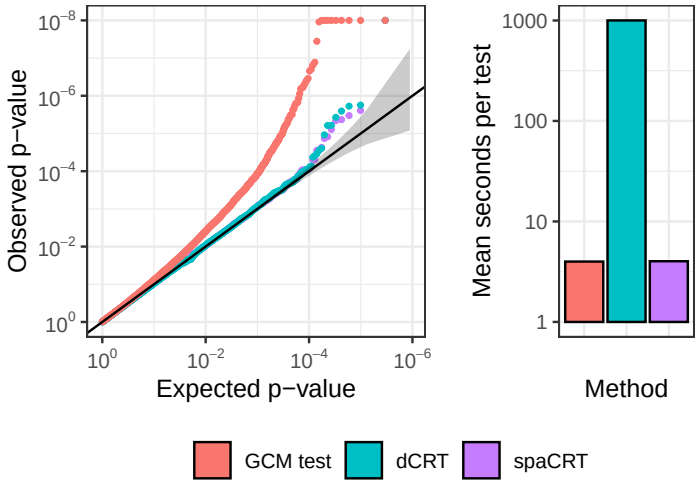

Figure 1: Comparing the Type-I error control and computation times the proposed spaCRT with other methods on the Gasperini et al. (2019) data. Left: QQ-plot of the p-values under the null hypothesis. Right: Mean computation times per hypothesis, in seconds.

1.4 Notation

Define $\text{sgn}(x)$ as the sign of x, i.e. $\text{sgn}(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 otherwise. For an infinitely differentiable function $f : \mathcal{X} \subset \mathbb{R} \mapsto \mathbb{R}$, define $f^{(r)}$ to be its r-th derivative. Define f', f'' as the first and second derivative of f respectively. Denote $[n]$ for any $n \in \mathbb{N}_+$ as $\{1, \ldots, n\}$. Define $\text{expit}(x) \equiv 1/(1 + \exp(-x))$. Define $\mathbb{E}_{\mathcal{L}_n}[\cdot \mid \mathcal{F}_n], \mathbb{E}_{\widehat{\mathcal{L}}_n}[\cdot \mid \mathcal{F}_n]$ as the conditional expectations under law $\mathcal{L}_n$ and its estimate $\widehat{\mathcal{L}}_n$ respectively. Similarly, we define $\text{Var}_{\mathcal{L}_n}[\cdot \mid \mathcal{F}_n]$ and $\text{Var}_{\widehat{\mathcal{L}}_n}[\cdot \mid \mathcal{F}_n]$ as the conditional variance counterpart. We use the following standard notations regarding the asymptotic properties of X_n:

$$
\begin{aligned}
&X_n = O_{\mathbb{P}}(1) && \text{if for each } \delta > 0 \text{ there is an } M > 0 \text{ s.t. } \limsup_{n\to\infty} \mathbb{P}[|X_n| > M] < \delta;\\
&X_n = \Omega_{\mathbb{P}}(1) && \text{if for each } \delta > 0 \text{ there is an } \eta > 0 \text{ s.t. } \limsup_{n\to\infty} \mathbb{P}[|X_n| < \eta] < \delta;\\
&X_n = o_{\mathbb{P}}(1) && \text{if } \mathbb{P}[|X_n| > \eta] \to 0 \text{ for all } \eta > 0.
\end{aligned}
$$

We use $a_n \sim b_n$ if $0 < \liminf_{n\to\infty} |a_n/b_n| \leq \limsup_{n\to\infty} |a_n/b_n| < \infty$ and $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| \to 0$.

2 The conditional saddlepoint approximation

Let $\{W_{in}\}_{1\leq i\leq n, n\geq 1}$ be a triangular array of random variables on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{F}_n \subseteq \mathcal{F}$ be a sequence of σ-algebras so that $\mathbb{E}[W_{in}|\mathcal{F}_n] = 0$ for each (i, n) and $\{W_{in}\}_{1\leq i\leq n}$ are independent conditionally on $\mathcal{F}_n$ for each n. For a sequence of cutoff values $w_n \in \mathcal{F}_n$, we will consider SPAs for the conditional tail probability:

$$
p_n \equiv \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right]. \tag{2}
$$

Conditional probability (2) arises naturally in p-value computation in the context of resampling-based hypothesis testing procedure. Consider a hypothesis test based on the statistic $T_n \equiv \frac{1}{n}\sum_{i=1}^n X_{in}$ and the resampling distribution $\widetilde{T}_n \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{X}_{in}$ for some resampling mechanism generating $\widetilde{X}_{in} \mid \sigma(X_{1n}, \ldots, X_{nn})$. Setting $W_{in} \equiv \widetilde{X}_{in}$, $w_n \equiv T_n$, and $\mathcal{F}_n \equiv \sigma(X_{1n}, \ldots, X_{nn})$, we find that the p-value of this test,

$$p_n \equiv \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^n X_{in} \;\middle|\; X_{1n}, \ldots, X_{nn}\right], \tag{3}$$

matches the definition (2). Thus an accurate analytic approximation to the conditional tail probability (2) is tempting to circumvent the necessity of repeated resampling. In this section, we will state our main results on the approximation accuracy of the saddlepoint approximation under two sets of tail assumptions on W_{in} (Section 2.1). We then discuss how the new results relate to the existing results in the literature (Section 2.2).

2.1 A new SPA for approximating conditional distribution

Now, we impose assumptions on the triangular array $\{W_{in}\}_{1\leq i\leq n, n\geq 1}$ in terms of their tail behaviors. In particular, we extend the definitions of sub-exponential and compactly supported random variables to the conditional setting and consider the *conditionally sub-exponential* (CSE) and *conditionally compactly supported* (CCS) conditions. We assume throughout that Assumption 1 or Assumption 2 holds.

Assumption 1 (CSE condition)**.** *There exist $\theta_n \in \mathcal{F}_n$ and $\beta > 0$ such that $\theta_n = O_{\mathbb{P}}(1)$ and almost surely,*

$$0 \leq \theta_n < \infty \quad \text{and} \quad \mathbb{P}[|W_{in}| \geq t \mid \mathcal{F}_n] \leq \theta_n \exp(-\beta t) \text{ for all } i, n \text{ and } t > 0.$$

Assumption 2 (CCS condition)**.** *There exist $\nu_{in} \in \mathcal{F}_n$ such that $\frac{1}{n}\sum_{i=1}^n \nu_{in}^4 = O_{\mathbb{P}}(1)$ and almost surely,*

$$0 \leq \nu_{in} < \infty \quad \text{and} \quad W_{in} \in [-\nu_{in}, \nu_{in}] \text{ for all } i \text{ and } n.$$

The saddlepoint approximation is usually based on the cumulant-generating functions (CGFs) of the summands. For our purposes, we use conditional CGFs:

$$K_{in}(s) \equiv \log \mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] \quad \text{and} \quad K_n(s) \equiv \frac{1}{n}\sum_{i=1}^n K_{in}(s). \tag{4}$$

The first step of the SPA is to find the solution $\hat{s}_n$ to the *saddlepoint equation*:

$$K_n'(s) = w_n. \tag{5}$$

We denote the solution to saddlepoint equation (5) as $\hat{s}_n$ and later Theorem 1 guarantees the existence and uniqueness of $\hat{s}_n$. With the solution $\hat{s}_n$ in hand, we define the saddlepoint approximation as follows. If we define

$$\lambda_n \equiv \hat{s}_n\sqrt{nK_n''(\hat{s}_n)}; \quad r_n \equiv \begin{cases} \operatorname{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_n w_n - K_n(\hat{s}_n))} & \text{if } \hat{s}_n w_n - K_n(\hat{s}_n) \geq 0; \\ \operatorname{sgn}(\hat{s}_n) & \text{otherwise,} \end{cases} \tag{6}$$

then the saddlepoint approximation is given by

$$\widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^n W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \tag{7}$$

When $w_n = 0$, we have $\hat{s}_n = \lambda_n = r_n = 0$. In this case, we take by convention that $1/0 - 1/0 \equiv 0$ in equation (9), so that $\widehat{\mathbb{P}}_{\mathrm{LR}} \equiv 1/2$. We will also use the convention $0/0 = 1$. The approximation $\widehat{\mathbb{P}}_{\mathrm{LR}}$ (7) is a direct generalization of the classical Lugannani-Rice formula (Lugannani and Rice, 1980) to the conditional setting.

Theorem 1. *Let W_{in} be a triangular array of random variables that are mean-zero and independent for each n, conditionally on $\mathcal{F}_n$. Suppose either Assumption 1 or Assumption 2 holds, and that*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[W_{in}^2 \mid \mathcal{F}_n] = \Omega_{\mathbb{P}}(1). \tag{8}$$

Let $w_n \in \mathcal{F}_n$ be a sequence with $w_n \xrightarrow{\mathbb{P}} 0$. Then, the saddlepoint equation (5) has a unique and finite solution $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$ with probability approaching 1 as $n \to \infty$. The explicit form of $\hat{s}_n$ can be found in Appendix B.1. Furthermore, the saddlepoint approximation $\widehat{\mathbb{P}}_{\mathrm{LR}}$ to the conditional tail probability (2) defined by equations (6) and (7) has vanishing relative error:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \tag{9}$$

We make several remarks on the resuls of Theorem 1.

Remark 1 (Generality and transparency of assumptions)**.** While most of the existing literature is fragmented based on whether the summands W_{in} are smooth or lattice, Theorem 1 (and later Corollary 1) unify these two cases by not making any such assumptions. The price we pay for this generality is that our guarantee (9) does not come with a rate, as compared to most existing results. Nevertheless, we find it useful to have a single result encompassing a broad range of settings. Our assumptions are not only general but also transparent. Most existing results require complicated and difficult-to-verify conditions, often stated in terms of transforms of W_{in}. By contrast, consider for example the Assuption 1 (sub-exponential summands) and condition (8) (non-degenerate variance). These are standard and easy to understand and verify.

Remark 2 (Justification of application to resampling-based tests)**.** The result (9) lays a solid mathematical foundation for applying the SPA to a range of resampling-based hypothesis tests with independently resampled summands, including the sign-flipping test, bootstrap-based tests, and the conditional randomization test (CRT). Even though our main focus in this paper is on CRT, result (9) provides **the first rigorous justification of an SPA for the sign-flipping test** of the kind originally proposed by Daniels (1955). We summarize such result in Theorem 5 in Appendix B.2.

Remark 3 (Technical challenges)**.** The proof of Theorem 1 presents several technical challenges, requiring us to significantly extend existing proof techniques. One of the most significant challenges is the conditioning, which adds an extra layer of randomness to the problem. For example, the saddlepoint equation (5) is a random equation, with random solution $\hat{s}_n$, which we must show exists with probability approaching 1. Furthermore, the cutoff w_n is random, and in particular we must handle cases

depending on the realization of the sign of this cutoff. A crucial step in our proof (which follows the general structure of that of Robinson, 1982) is to use the Berry-Esseen inequality, but the extra conditioning requires us to prove a new conditional Berry-Esseen theorem. Another challenge is that we allow w_n to decay to zero at an arbitrary rate, which requires a delicate analysis of the convergence of the SPA formula appearing in the RHS of the result (9).

Remark 4 (Assumptions on the threshold w_n)**.** The assumption $w_n = o_{\mathbb{P}}(1)$ in Theorem 1 appears to be restrictive at the first glance. However, this assumption allows us to guarantee the existence of a solution the saddlepoint equation beyond the case of compact support (Daniels, 1954). On the other hand, the rate of convergence of w_n towards zero can be arbitrary, accommodating for two statistically meaningful regimes: (a) Moderate deviation regime: $w_n = O_{\mathbb{P}}(n^{-\alpha}), \alpha \in (0, 1/2)$ and (b) CLT regime: $w_n = O_{\mathbb{P}}(1/\sqrt{n})$. We discuss this assumption further in Remark 5 after Theorem 5.

2.2 Connection to existing unconditional results

It turns out our results are closely connected to several existing results in the literature. First, Theorem 1 reduces to the following variant of the classical Lugannani and Rice (1980) result by setting $\mathcal{F}_n \equiv \{\varnothing, \Omega\}$:

Corollary 1. *Let W_{in} be a triangular array of random variables that are mean-zero and independent for each n. Suppose that each W_{in} is sub-exponential with constants $\theta, \beta > 0$, i.e.*

$$\mathbb{P}[|W_{in}| \geq t] \leq \theta \exp(-\beta t) \quad \text{for all } t > 0. \tag{10}$$

Furthermore, suppose that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[W_{in}^2] > 0. \tag{11}$$

Given a sequence of cutoffs $w_n \to 0$ as $n \to \infty$, the saddlepoint equation (5) (based on the unconditional CGF K_n) has a unique solution $\hat{s}_n$ on $[-\varepsilon/2, \varepsilon/2]$ for all sufficiently large n. Furthermore, the unconditional saddlepoint approximation $\widehat{\mathbb{P}}_{\mathrm{LR}}$ (based on the unconditional CGF K_n) has vanishing relative error

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right](1 + o(1)). \tag{12}$$

Compared to the classical results, the significance of Theorem 1 and Corollary 1 is the generality to underlying distributions and transparency on the assumptions as discussed in Remark 1. Aside from the Lugannani-Rice formula, another tail probability estimate is proposed in Robinson, 1982. In fact, we present an extension of Theorem 1 in Proposition 1 that employs a conditional variant of Robinson's formula:

$$\widehat{\mathbb{P}}_{\mathrm{R}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] \equiv \exp\left(\frac{\lambda_n^2 - r_n^2}{2}\right)(1 - \Phi(\lambda_n)). \tag{13}$$

Proposition 1. *Under the assumptions of Theorem 1,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{R}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \tag{14}$$

In other words, $\widehat{\mathbb{P}}_{\mathrm{LR}}$ (7) is equivalent to $\widehat{\mathbb{P}}_{\mathrm{R}}$ (13) with relative error $o_{\mathbb{P}}(1)$, linking Robinson's formula to that of Lugannani and Rice. A similar result was proved in the unconditional case by Kolassa (2007).

3 spaCRT: A resampling-free approximation to dCRT

3.1 Background: dCRT and GCM test

The dCRT procedure, as proposed by Liu et al. (2022), is designed under the *model-X assumption* that $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is known (Candès et al., 2018). However, this procedure is usually deployed in practice by learning this conditional distribution in-sample. In a prior work, we established the statistical properties of the dCRT with $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ estimated in sample (Niu et al., 2024). In this paper, we will refer to the latter procedure as the dCRT, allowing a minor abuse of terminology. Furthermore, we consider the special but still fairly general case when

$$\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z}) = f(\boldsymbol{X} \mid \theta_{n,x}(\boldsymbol{Z})), \tag{15}$$

where $f(x|\theta)$ is an exponential family with natural parameter θ, natural parameter space $\mathbb{R}$, and log-partition function A:

$$f(x|\theta) = \exp(\theta x - A(\theta))h(x). \tag{16}$$

This is not a restrictive assumption, since we allow the function $\theta_{n,x}(\boldsymbol{Z})$ to be arbitrary. Given this setup, consider estimating the functions $\theta_{n,x}(\boldsymbol{Z})$ and $\mu_{n,y}(\boldsymbol{Z}) \equiv \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y} \mid \boldsymbol{Z}]$ by $\widehat{\theta}_{n,x}(\boldsymbol{Z})$ and $\widehat{\mu}_{n,y}(\boldsymbol{Z})$, respectively. The learning procedures for these quantities can be arbitrary. The choices include but are not limited to parametric, non-parametric and high-dimensional regression methods. Setting $\widehat{\mu}_{n,x}(\boldsymbol{Z}) \equiv A'(\widehat{\theta}_{n,x}(\boldsymbol{Z}))$, we arrive at the test statistic

$$T_n^{\mathrm{dCRT}}(X, Y, Z) = \frac{1}{n}\sum_{i=1}^{n}(X_{in} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})). \tag{17}$$

The dCRT is obtained by comparing $T_n^{\mathrm{dCRT}}(X, Y, Z)$ to a null distribution obtained by resampling $X_{in} \mid Z_{in}$ based on the estimated distribution $f(\cdot \mid \widehat{\theta}_{n,x}(Z_{in}))$. This procedure is summarized in Algorithm 1.

Algorithm 1: dCRT **procedure with exponential family for** $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$

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

Learn $\widehat{\theta}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,x}(\cdot)$ based on (X, Z); learn $\widehat{\mu}_{n,y}(\cdot)$ based on (Y, Z);

Compute $T_n^{\mathrm{dCRT}}(X, Y, Z)$ as in (17);

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

  Sample $\widetilde{X}^{(m)}|X, Y, Z \sim \prod_{i=1}^{n} f(\cdot \mid \widehat{\theta}_{n,x}(Z_{in}))$ and compute

$$T_n^{\mathrm{dCRT}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{n}\sum_{i=1}^{n}(\widetilde{X}_{in}^{(m)} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})); \tag{18}$$

end

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

Unlike dCRT procedure, *generalized covariance measure* (GCM) test uses the asymptotic distribution of the same test statistic (17) to find the p-value. Therefore, the computation of the p-value is much faster than that of dCRT because it does not require repeated resampling. However, the resampling nature of dCRT (Algorithm 1), though computationally intensive, can lead to improved finite-sample performance. Recall Figure 1. The key insight is that the GCM test, which relies on asymptotic normality, can suffer from slow convergence in the presence of excessive sparsity in the data. In contrast, dCRT leverages resampling, and the resulting distribution of the resampled test statistic can more accurately approximate the sampling distribution of the observed statistic. Such intuition can be mathematically justfied in an idealized setup and we show how GCM can suffer from the slow convergence rate of Type-I error while dCRT preserves the Type-I error control in Theorem 6 in Appendix C.2.

We refer reader to Appendix C for more details on the comparison between GCM and dCRT and their common features such as double robustness (Appendix C.1).

3.2 The spaCRT

The resampling procedure in Algorithm 1 is slow and can be infeasible for large datasets. In this section, we propose a new test, the *saddlepoint approximation-based conditional randomization test* (spaCRT), which is completely resampling-free and has similar finite-sample performance as dCRT. If we consider the limit of the dCRT p-value as the number of resamples M grows indefinitely, we obtain

$$p_{\text{dCRT}} \equiv \mathbb{P}\left[T_n^{\text{dCRT}}(\widetilde{X}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z) \mid X, Y, Z\right]. \tag{19}$$

We approximate this conditional tail probability via the SPA. Note that the resampled test statistic defined in (18) is the mean of conditionally independent random variables:

$$T_n^{\text{dCRT}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{n}\sum_{i=1}^{n} W_{in},\ W_{in} \equiv a_{in}(\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})),\ a_{in} \equiv Y_{in} - \widehat{\mu}_{n,y}(Z_{in}).$$

Indeed, W_{in} are independent, but not identically distributed, conditionally on the σ-algebra $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. Fitting the formulation to (2), we will use $K_{in}(s|\mathcal{F}_n) \equiv K_{in}(s)$ and $K_n(s|\mathcal{F}_n) \equiv K_n(s)$ as defined in (4). The closed-form expressions for the $K_n(s|\mathcal{F}_n)$ and its derivatives under natural exponential family (16) are available in Appendix E.

Algorithm 2: spaCRT **procedure**

Input: Data (X, Y, Z).

1 Learn $\widehat{\theta}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,x}(\cdot)$ based on (X, Z), $\widehat{\mu}_{n,y}(\cdot)$ based on (Y, Z);

2 Compute $T_n^{\mathrm{dCRT}}(X, Y, Z)$ as in (17);

3 Find $\hat{s}_n$ that solves the saddlepoint equation

$$K_n'(s \mid \mathcal{F}_n) = T_n^{\mathrm{dCRT}}(X, Y, Z); \tag{20}$$

4 Compute $\lambda_n = \sqrt{n}\hat{s}_n\sqrt{K_n''(\hat{s}_n \mid \mathcal{F}_n)}$ and

$$r_n = \begin{cases} \mathrm{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_n T_n^{\mathrm{dCRT}} - K_n(\hat{s}_n \mid \mathcal{F}_n))}, & \text{if } \hat{s}_n T_n^{\mathrm{dCRT}} - K_n(\hat{s}_n \mid \mathcal{F}_n) \geq 0; \\ \mathrm{sgn}(\hat{s}_n) & \text{otherwise.} \end{cases}$$

Output: spaCRT p-value

$$p_{\mathrm{spaCRT}} \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \tag{21}$$

Then we can apply the result in Theorem 1 to obtain a closed-form approximation towards dCRT p-value (19). To this end, we present the spaCRT procedure (Algorithm 2). The spaCRT procedure is attractive because it is completely resampling-free. It requires the following one-time computations: fitting the estimates $\widehat{\theta}_{n,x}$ and $\widehat{\mu}_{n,y}$, calculating the test statistic T_n^{dCRT}, and finding the solution to the saddlepoint equation (20). The latter is a one-dimensional root-finding problem and can be solved efficiently using standard numerical optimization algorithms.

To make the spaCRT procedure more concrete, we provide an example in the case that $\boldsymbol{X}$ is binary, a setting that matches our motivating application.

Example 1 (Bernoulli sampling)**.** Suppose $\boldsymbol{X} \mid \boldsymbol{Z} \sim \mathrm{Ber}(\mu_{n,x}(\boldsymbol{Z}))$, and $\theta_{n,x}(\boldsymbol{Z}) = \mathrm{logit}(\mu_{n,x}(\boldsymbol{Z}))$. Then, we have $A(\theta) = \log(1 + \exp(\theta))$. After some manipulation, the saddlepoint equation reduces to

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))(X_{in} - \mathrm{expit}(\widehat{\theta}_{n,x}(Z_{in}) + s(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})))) = 0.$$

Defining $\widetilde{\mu}_{n,x}(Z_{in}) \equiv \mathrm{expit}(\widehat{\theta}_{n,x}(Z_{in}) + \hat{s}_n(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})))$ for convenience, λ_n and r_n can be computed as $\lambda_n = \hat{s}_n\sqrt{\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2\widetilde{\mu}_{n,x}(Z_{in})(1 - \widetilde{\mu}_{n,x}(Z_{in}))}$ and

$$r_n = \mathrm{sgn}(\hat{s}_n)\sqrt{2\sum_{i=1}^{n}\left(X_{in}\log\frac{\widetilde{\mu}_{n,x}(Z_{in})}{\widehat{\mu}_{n,x}(Z_{in})} + (1 - X_{in})\log\frac{1 - \widetilde{\mu}_{n,x}(Z_{in})}{1 - \widehat{\mu}_{n,x}(Z_{in})}\right)},$$

or simply $\mathrm{sgn}(\hat{s}_n)$ if the quantity under the square root is negative. Putting these pieces together, the spaCRT p-value can be computed as in equation (21).

4 Theoretical results of the spaCRT

spaCRT does not require any resampling and thus has a significant advantage over dCRT in terms of computation. This advantage does not come with a sacrifice of

statistical accuracy. In this section, we establish the theoretical properties of the spaCRT.

4.1 General theory: approximation and Type-I errors

We first state general results concerning the approximation accuracy of the spaCRT p-value and then prove the asymptotic Type-I error control of spaCRT.

Theorem 2 (Approximation accuracy)**.** *Suppose there exists $S > 0$ such that **one** of the following conditions holds:*

$$\sup_i |\widehat{\theta}_{n,x}(Z_{in})|,\ \sup_i |\widehat{\mu}_{n,y}(Z_{in})| = O_{\mathbb{P}}(1),\ \mathbb{P}[Y_{in} \in [-S, S]] = 1 \text{ for any } i, n; \tag{CSE}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1),\ \mathbb{P}\left[\widetilde{X}_{in} \in [-S, S]\right] = 1 \text{ for any } i, n. \tag{CCS}$$

Suppose the following conditions hold:

$$|\widehat{\theta}_{n,x}(Z_{in})| < \infty, |\widehat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i, n \text{ almost surely}; \tag{22}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 A''(\widehat{\theta}_{n,x}(Z_{in})) = \Omega_{\mathbb{P}}(1); \tag{23}$$

$$T_n^{\mathrm{dCRT}}(X, Y, Z) \xrightarrow{\mathbb{P}} 0. \tag{24}$$

Then, the saddlepoint equation (20) has a unique and finite solution $\hat{s}_n \in [-1/16, 1/16]$ with probability approaching 1 as $n \to \infty$. Furthermore, $\mathbb{P}[p_{\mathrm{spaCRT}} > 0] \to 1$ as $n \to \infty$ and the spaCRT p-value p_{spaCRT} approximates the dCRT p-value p_{dCRT} with vanishing relative error:

$$p_{\mathrm{dCRT}} = p_{\mathrm{spaCRT}} \cdot (1 + o_{\mathbb{P}}(1)) \tag{25}$$

To better understand the assumptions in Theorem 2, we provide the following remarks.

Remark 5 (Comment on condition (24))**.** The role of the condition (24) is to guarantee the existence of the solution to the saddlepoint equation. This assumption allows the test statistic $T_n^{\mathrm{dCRT}}(X, Y, Z)$ to converge to zero in probability, **at any rate**. In particular, under the null hypothesis and contiguous local alternatives, this condition holds.

Remark 6 (Relative error guarantee)**.** The relative error guarantee in conclusion (25) is a strong result and is a direct consequence of result (9). It means not only the difference of p-values is close to 0 with probability approaching 1, but also the ratio of p-values is close to 1 with probability approaching 1. This is a particularly desirable property for approximating small p-values.

Defining the level-α tests associated with the dCRT and spaCRT p-values:

$$\phi_{n,\alpha}^{\mathrm{dCRT}} \equiv \mathbb{1}(p_{\mathrm{dCRT}} \le \alpha) \quad \text{and} \quad \phi_{n,\alpha}^{\mathrm{spaCRT}} \equiv \mathbb{1}\left(p_{\mathrm{spaCRT}} \le \alpha\right),$$

the following theorem states that the asymptotic Type-I error control of the spaCRT follows from that of the dCRT.

Corollary 2 (Asymptotic validity of spaCRT)**.** *Suppose the assumptions of Theorem 2 hold. Fix* $\alpha \in (0,1)$*. If* $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$*, then we have*

$$\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \leq \alpha] \leq \alpha.$$

Corollary 2 states the asymptotic validity of spaCRT given the asymptotic validity of dCRT. dCRT is proposed originally assuming the exact knowledge of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$, or the so-called Model-X assumption. Under the model-X assumptions, the p-value produced by dCRT procedure is exact, i.e., $\mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$. Therefore, p-value obtained from spaCRT procedure is asymptotically valid under the assumptions of Theorem 2. For asymptotic validity of dCRT with general in-sample fit $\widehat{\mathcal{L}}_n(\boldsymbol{X} \mid \boldsymbol{Z})$, we refer reader to the detailed discussion in Niu et al. (2024).

4.2 Case studies with modern regression techniques

We will dedicate this section to special cases to show approximation accuracy of spaCRT and its validity under null hypothesis. In particular, we investigate the performance of spaCRT when $\boldsymbol{Y}|\boldsymbol{Z}$ is estimated using modern regression techniques. We will consider low-dimensional and high-dimensional generalized linear regressions in the main text and we postpone the discussion of the nonparametric regression with kernel ridge regression to Appendix M.1. We need two assumptions to state the results.

Assumption 3. $0 < \inf_n \mathbb{E}[(X_{in} - \mathbb{E}[X_{in} \mid Z_{in}])^2 (Y_{in} - \mathbb{E}[Y_{in} \mid Z_{in}])^2]$.

Assumption 4. *Support of* $\boldsymbol{Z} \in \mathbb{R}^d$ *is compact, i.e.,* $\|\boldsymbol{Z}\|_\infty \leq C_Z$ *for* $C_Z \in (0, \infty)$.

Throughout this section, we will consider the following logistic model for $\boldsymbol{X}|\boldsymbol{Z}$:

$$\boldsymbol{X}|\boldsymbol{Z} \sim \text{Ber}(\text{expit}(\boldsymbol{Z}^\top \theta_n)).$$

The choice of logistic model is mainly for the ease of the theoretical analysis. In fact, the spaCRT can be easily integrated with diverse fitting methods beyond just logistic regression under binary valued $\boldsymbol{X}$, for example hidden Markov models (in Section 5.2) and nonparametric random forest classification (in Appendix M.2).

Low-dimensional generalized linear regression. Suppose we are under the classical low-dimensional setup so that we write $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$ and $\theta_n = \theta$. To echo Figure 1, we first consider the conditional distribution $\boldsymbol{Y} \mid \boldsymbol{Z}$ to be $\mathcal{P}_y(A_y'(\boldsymbol{Z}^\top \beta))$ where $\mathcal{P}_y$ is a distribution in the natural exponential family (16) and A_y is the log-parition functions for $\boldsymbol{Y}|\boldsymbol{Z}$. Now we state our result.

Theorem 3. *Consider* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$*. Suppose assumptions 3-4 hold. Then if the maximum likelihood estimates (MLEs)* $\widehat{\beta}, \widehat{\theta}$ *satisfy*

$$\|\widehat{\beta} - \beta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}) \quad \text{and} \quad \|\widehat{\theta} - \theta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}), \tag{26}$$

then the conclusion in Theorem 2 holds and we have $\phi_{n,\alpha}^{\text{spaCRT}}$ *is asymptotically valid, i.e.* $\lim_{n\to\infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$.

We want to emphasize the $\sqrt{n}$ rate condition in (26) is classical for MLEs under mild regularity conditions.

High-dimensional regression. We establish the validity of the spaCRT under the following generalized linear models for $\boldsymbol{Y}|\boldsymbol{Z}$ $\boldsymbol{Y}|\boldsymbol{Z} \sim \mathcal{P}_y(A'_y(\boldsymbol{Z}^\top \beta_n))$ and this setup is similar to the low-dimension case in the previous part while here we allow dimension of $\boldsymbol{Z}$ (as well as β_n) to grow with sample size n. We will demonstrate how spaCRT can be used in the presence of high-dimensional parameters. In particular, we consider the estimators $\widehat{\beta}_n, \widehat{\theta}_n$ for β_n, θ_n are obtained from the lasso estimators (Tibshirani, 1996). The definitions of these estimators are standard and can be found in Appendix H.4. Recovering the true parameters β_n, θ_n is challenging (sometimes even unidentifiable) in the high-dimensional setting unless certain structure assumptions on these parameters are imposed. We show that the spaCRT is valid under the sparse signal setup. We start by stating the following assumption on the covariate distribution $\boldsymbol{Z}$.

Assumption 5 (Design assumption)**.** *Suppose the distribution of* $\boldsymbol{Z}$ *satisfies*

$$\text{minimum eigenvalue of } \mathbb{E}[\boldsymbol{Z}\boldsymbol{Z}^\top] \text{ is boundeed away from } 0; \tag{27}$$

$$\mathbb{E}[\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4] \leq \kappa < \infty, \ \forall \eta \in \mathbb{R}^d, \|\boldsymbol{\eta}\|_2 = 1. \tag{28}$$

Theorem 4. *Consider* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$*. Suppose assumptions 3-5 hold and* $\exists \delta \in (0,1)$*,*

$$\max\{1, s_{\theta_n}, s_{\beta_n}\}\sqrt{\log(d)/n} \sim n^{-\delta} \text{ where } (s_{\beta_n}, s_{\theta_n}) \equiv (\|\beta_n\|_0, \|\theta_n\|_0); \tag{29}$$

$$\sup_n \|\theta_n\|_1 < \infty, \ \sup_n \|\beta_n\|_1 < \infty. \tag{30}$$

Then if we choose $\lambda_n = C_\lambda\sqrt{\log(d)/n}$ *and* $\nu_n = C_\nu\sqrt{\log(d)/n}$ *for some universal constants* C_λ, C_ν*, the conclusion in Theorem 2 hold. If additionaly, we have*

$$s_{\theta_n} \cdot s_{\beta_n} \cdot \frac{\log(d)}{n^{1/2}} = o(1) \tag{31}$$

then $\phi^{\text{spaCRT}}_{n,\alpha}$ *is asymptotically valid, i.e.* $\lim_{n\to\infty} \mathbb{E}[\phi^{\text{spaCRT}}_{n,\alpha}] = \alpha$ *for any* $\alpha \in (0,1)$*.*

Let us comment on the assumptions required. Assumption 5 imposes the structures on the covariate distribution which are commonly required in the high-dimensional regression. In fact, conditions (27)-(28) are also required in Example 9.17 and Theorem 9.36 in Wainwright (2019) when proving the *restricted strong convexity*, a fundamental property involved when proving consistency results of high-dimensional regression coefficients (see for example Corollary 9.26 in Wainwright (2019)). Condition (29) regulates the rate growth between $s_{\theta_n}, s_{\beta_n}, d$ and n. The boundedness condition (30) is a relatively mild condition required by showing the almost sure convergence of $\widehat{\mu}_{n,y}(\cdot), \widehat{\theta}_{n,x}(\cdot)$ when verifying condition (22).

5 Application to data with excessive sparsity

As discussed in Section 1, the spaCRT method can be very useful when it comes to analysis with very sparse data, i.e. excessive zeros in outcome data Y and/or in treatment X. In the following sections, we evaluate the finite-sample performances of spaCRT in two simulation setups inspired by single-cell CRISPR screens analysis and GWAS with rare disease and rare genetic variant, respectively.

5.1 Single-cell CRISPR screens analysis

Simulation setup: To mimic the application of single-cell CRISPR screens, we model $\boldsymbol{X} \mid \boldsymbol{Z}$ as a logistic regression model and $\boldsymbol{Y} \mid \boldsymbol{Z}$ as a negative binomial regression model (Barry et al., 2024; Barry et al., 2021; Gasperini et al., 2019). The latter modeling choice is quite common not just in single-cell CRISPR screen analysis but in single-cell RNA-seq analysis more broadly (Huang et al., 2018; Townes et al., 2019; Svensson, 2020). We consider the following data-generating model:

$$\boldsymbol{Z} \sim N(0,1);\ \boldsymbol{X} \mid \boldsymbol{Z} \sim \mathrm{Ber}(\mathrm{expit}(\gamma_0 + \boldsymbol{Z}));\ \boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z} \sim \mathrm{NB}(\exp(\beta_0 + \rho \boldsymbol{X} + \boldsymbol{Z}), r), \quad (32)$$

where $r > 0$ is the *size parameter* controlling the overdispersion of the negative binomial distribution. Here, $\boldsymbol{X}$, $\boldsymbol{Y}$, and $\boldsymbol{Z}$ represent the indicator of perturbation presence, gene expression, and a single covariate with a confounding effect, respectively. The parameters γ_0 and β_0 control the proportion of cells with perturbations and the mean expression of the gene, respectively, and therefore control the sparsity level of X and Y. The source for the sparsity from real data can be found in Appendix K.1. In this setup, we consider testing for association between a single CRISPR perturbation and a single gene. The parameter ρ controls the strength of the signal, i.e., the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ conditional on $\boldsymbol{Z}$. Therefore, $\rho = 0$ and $\rho \neq 0$ corresponds to the null and alternative hypotheses, respectively.

Methodologies compared: We compare the four tests: spaCRT, dCRT, GCM test and score test. The detailed discussion of the methods can be found in Appendix K.2. Note we use $M = 10,000$ resamples for dCRT for the accuracy of p-value.

Simulation results Here, we present a representative selection of simulation results (Figure 2). These results correspond to $r = 0.05$ and $\beta_0 = -5$, and all tests are applied at nominal level $\alpha = 0.01$. Additional results are provided in Appendix K.3. We find from Figure 2a, which displays p-value distributions under the null hypothesis, that the spaCRT and dCRT tests have similar p-value distributions, both of which are close to uniform. Meanwhile, the GCM test behaves too liberally for left-sided tests and too conservatively for right-sided tests, while the score test behaves too conservatively for left-sided tests and too liberally for right-sided tests. These trends are reflected in the Type-I error rates and powers in Figure 2b,c. We remark that the spaCRT and dCRT tests control Type-I error for all settings of γ_0, though both tests tend to become conservative as X becomes sparser. Furthermore, the spaCRT and dCRT are the most powerful tests among those that have Type-I error control for every parameter setting.

Next, we remark on how the methods' performance is impacted by the problem parameters γ_0 and β_0. As either X or Y becomes less sparse (i.e., as γ_0 or β_0 increase), the p-value distributions, Type-I error rates, and powers for the GCM and score tests improve. This is to be expected, as the test statistics converge more quickly towards the standard normal distribution when the quantities being averaged are less sparse. We defer the disucssion on size parameter r to Appendix K.3 where we will show dCRT and spaCRT are more robust to the values of r than the GCM and score tests.

Moreover, Figure 2d displays the computing times for the different methods in the settings considered in Figure 2a,b,c. We see that the spaCRT and GCM test are

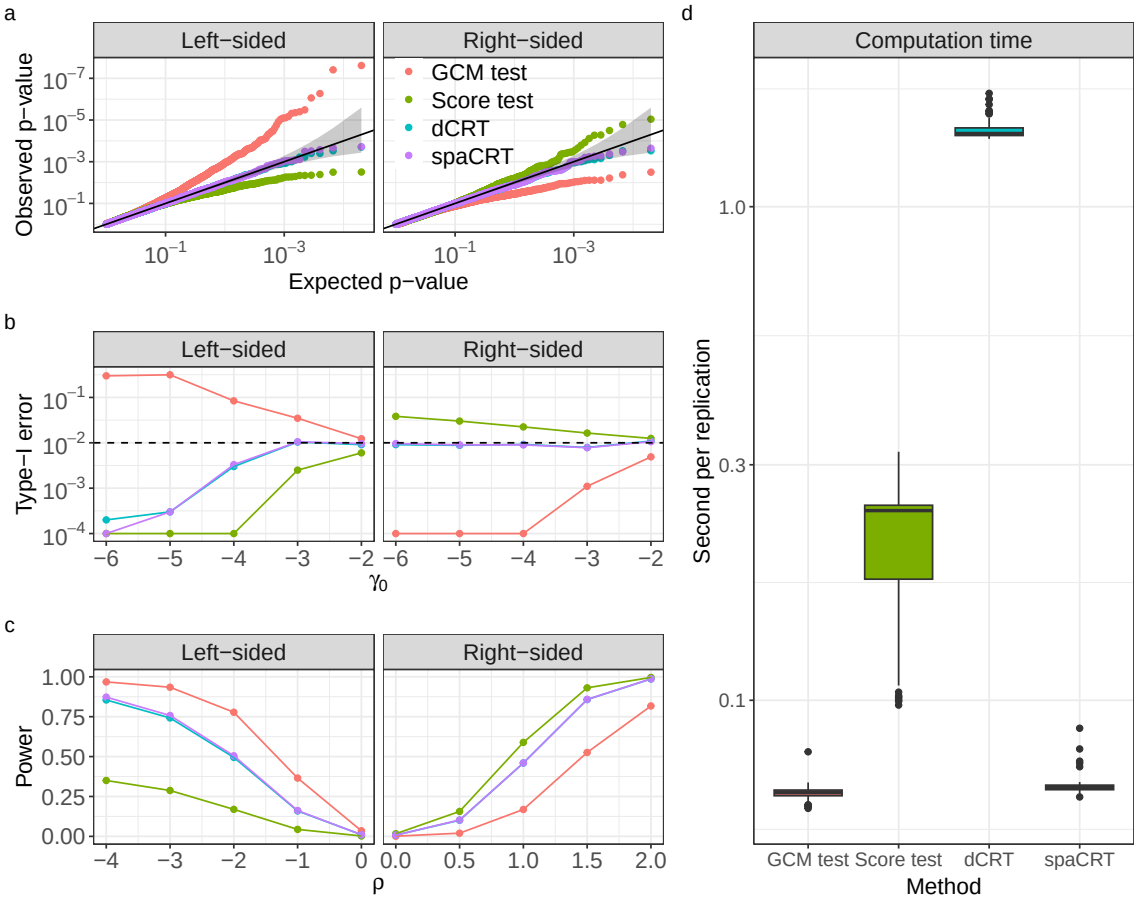

Figure 2: Summary of CRISPR screens simulation results for size parameter $r = 0.05$ and the significance level is set to 0.01. (a) QQ-plots of the p-values obtained under the null hypothesis for $(\gamma_0, \beta_0) = (-3, -5)$. (b) Type-I error rates for $\beta_0 = -5$ as a function of the sparsity of X (γ_0). (c) Power for $(\gamma_0, \beta_0) = (-3, -5)$ as a function of the signal strength (ρ). (d) The boxplot shows the time consumed across different simulation parameters while each point is the averaged time consumption over $10,000$ replications.

roughly tied for fastest, the score test is roughly half an order of magnitude slower than these two, while the dCRT is more than an order of magnitude slower.

Even though the CRISPR screens analysis is often concerned about multiple genes and perturbations, our simulation setup already captures the relevant statistical phenomena on the single perturbation-gene pair. To complete our narrative, we treat the p-values obtained from the simulation setup as arising under a multiple testing regime, and report the number of rejections after applying multiplicity corrections in Appendix K.3. These results are consistent with the simulations presented in Figure 6.

5.2 GWASs with rare diseases and genetic variants

Simulation setup: The goal of identifying genetic variaions associated with phenotypes in GWAS can be formulated as a conditional independence testing problem

(Sesia, Sabatti, and Candès, 2019):

$$H_0^j : \boldsymbol{X}_j \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{X}_{-j} \quad j = 1, \ldots, d \tag{33}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)^\top \in \{0,1\}^d$ is typically referring to the presence or absence of a specific allele at a given locus and $\boldsymbol{Y}$ is the phenotype of interest. Often time, a large scale loci are considered simultaneously. In other words, this is a high-dimensional setup and we consider $\boldsymbol{Z} \equiv \boldsymbol{X}_{-j}$ as covariate. We consider a binary outcome $\boldsymbol{Y}$, which may be viewed as indicator of the presence of a disease:

$$\boldsymbol{Y} \mid \boldsymbol{X} \sim \text{Ber}(\text{expit}(\gamma_0 + \boldsymbol{X}^\top \beta)).$$

Then under this model, the variable selection problem (33) is now euqivalent to testing $\beta_j = 0$ or not (Candès et al., 2018). Similarly, γ_0 here controls the sparsity of outcome Y and we consider $\{-3, -2\}$ for *high* and *low* sparsity setting, mimicking the relative frequency of phenotype occurrence. More explanation on the source of sparsity can be found in Appendix L.2.

We consider the genetic variable $\boldsymbol{X} \in \{0,1\}^d$ is generated from a *hidden Markov model* (HMM). HMMs have been broadly adopted to describe haplotypes, the sequence of alleles at a series of markers along one chromosome, and offer a good phenomenological description of the dependence between the explanatory variables in GWAS (Scheet and Stephens, 2006; Marchini et al., 2007; Browning and Browning, 2007). A brief introduction of the HMM can be found in Appendix J. We consider two sets of parameters generating HMMs with one setup mimicking the rare genetic variation scenario (high sparsity in X) and the other setup mimicking the common genetic variation scenario (low sparsity in X).

We will consider $d = 500$ and sample size $n = 2000$ in the simulation and leave the other details on parameters used in the simulation to Appendix K.2.

Methodologies compared: We consider the four tests: spaCRT, dCRT, GCM test and the Knockoff procedure (Barber and Candès, 2015). Methods spaCRT, dCRT and GCM use the same fitting methods and only differ on their calibration approach for the p-values. The method details can be found in Appendix L. We apply a both-sided test for all four methods. For spaCRT, dCRT and GCM, the p-values are corrected for multiple testing using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The FDR is controlled at level 0.1.

Simulation results We present a representative selection of simulation results (Figure 3) with high sparsity in X and the other results can be found in Appendix L.4. Note that the Knockoff procedure does not provide p-values but just a rejection set. Thus controlling family-wise error rate (FWER) is more challenging for the Knockoff procedure due to the lack of p-values. Therefore we only present the FDR plot (Figure 3a) and power plot (Figure 3b) in this simulation setup. We find from Figure 3a that the spaCRT and dCRT tests have similar FDR, both of which are close to the nominal level. Meanwhile, the GCM test behaves too liberally due to excessive sparsity in X and Y although the degree of inflation is more severe when $\gamma_0 = -3$. The Knockoff method can also control FDR while being conservative when signal is weak.

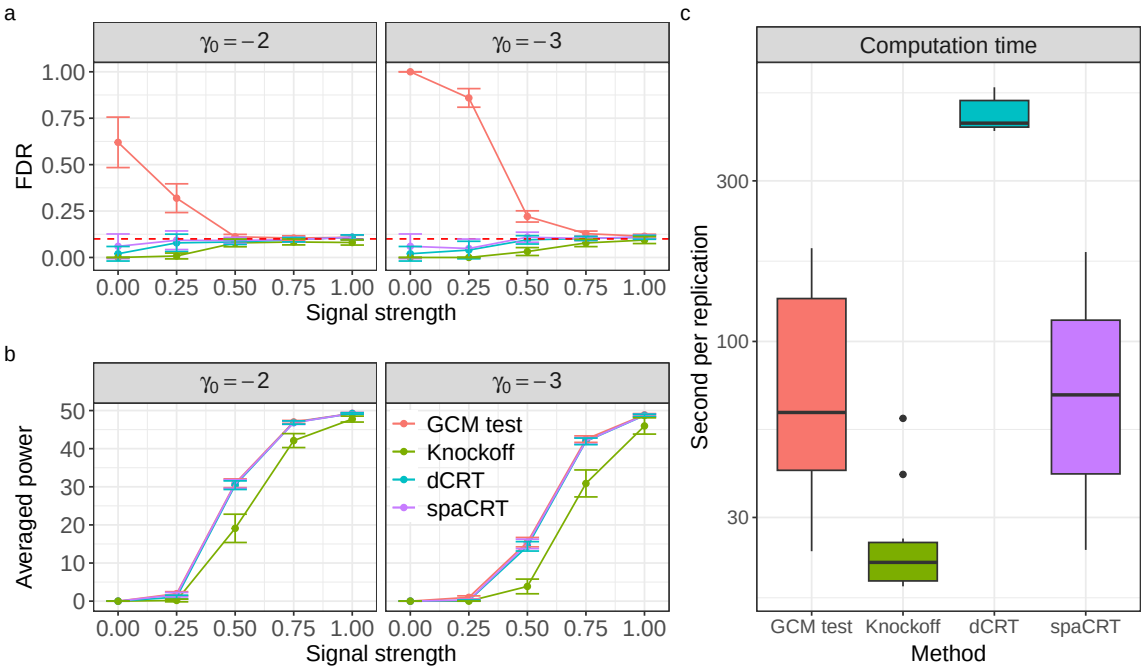

Figure 3: Summary of numerical simulation results GWAS with high sparsity in data X. All the results are obtained with the regularization parameter $\lambda =$ `lambda.1se`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Time consumed by different methods across 50 replications and all the varying parameters.

Figure 3b shows GCM, dCRT and spaCRT have similar power whereas Knockoff procedure tends to be less powerful. We believe the comparably low power in Knockoff may be due to sensitivity of Knockoff procedure to the distribution of $\boldsymbol{X}$, which can heavily affect the construction of decorrleated knockoff variable $\tilde{\boldsymbol{X}}$ to observed variable $\boldsymbol{X}$. Therefore we are reluctant to claim that Knockoff procedure is generally less powerful but postpone the more comprehensive investigation to future work.

Figure 3c shows the time consumption for different methods. Among all methods, Knockoff is the computationally most efficient method and this hinges on the fact that only one high-dimensional regression fit of $\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{X}}$ is required, where here $\tilde{\boldsymbol{X}}$ is the constructed knockoff variable. On the other hand, the other three methods involve d regressions for $\boldsymbol{Y}|\boldsymbol{X}_{-j}, j \in [d]$ if no further acceleration is applied, not to mention the excessive the resampling required in dCRT. With this being said, we employ a *tower trick* to boost the computation of $\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X}_{-j}]$ for dCRT, GCM and spaCRT. The implementation details can be found in Appendix L.3. The acceleration can push the computation time of spaCRT and GCM to be within a factor of 2 to 3 of the Knockoff procedure while being much faster than the dCRT. Note we use `R` package `SNPknock` to generate knockoff variable and the construction procedure is implemented in `Rcpp` for speed. Currently, spaCRT is implemented mostly in `R` so we expect further speedup when employing more efficient implementation.

6 Real data analysis

In this section, we compare the performance of the spaCRT to those of alternative methods on the analysis of the Gasperini et al. (2019) single-cell CRISPR screen dataset. We refer reader to Appendix N.1 for detailed description of the dataset.

6.1 Analyses conducted

Hypotheses tested. In order to assess the Type-I error and power of the methods compared, we will use CRISPR perturbations intended as negative (under null) and positive controls (under alternative), respectively. In particular, for Type-I error analysis, we test for association between each of the 51 negative control perturbations and each of 3,000 randomly sampled genes, for a total of $51 \times 3{,}000 = 153{,}000$ tests. We subsample the genes to reduce the computational burden of the analysis. To assess the power of each method, we test for association between each of the 754 positive control perturbations targeting genes and the gene they target, for a total of 754 tests.

Methods compared. We compare essentially the same methods as in the numerical simulations (recall Section 5.1). The only difference is that we replace the dCRT with a faster variant implemented in the R package `sceptre` (Barry et al., 2024; Barry et al., 2021), in order to make the analysis computationally feasible. The `sceptre` implementation of the dCRT fits a parametric curve to the resampling distribution of the test statistic based on a smaller number of resamples (a heuristic acceleration that is not theoretically justified). Furthermore, it is implemented in C++ for speed, unlike the other methods we consider, which are implemented in R. We apply left- and right-sided variants of each test on the negative control perturbation-gene pairs. For the positive control pairs, we apply only left-sided tests, since we are testing for a perturbation-induced decrease in gene expression.

6.2 Results

Type-I error. Figure 4a displays QQ plots of the negative control p-values obtained from all four methods. The two tests relying on asymptotic normality, the GCM and score tests, exhibit severe p-value inflation for left- and right-sided tests, respectively. This finding is consistent with our simulation results (Figure 2). On the other hand, the spaCRT and `sceptre` tests control Type-I error well for both left- and right-sided tests. We also report the number of false discoveries on the negative control pairs in Appendix N.2; the message from these results are consistent with that of the single testing results.

Next, we investigate the impact of the problem sparsity on calibration. Following (Barry et al., 2024), we measure sparsity in terms of the *effective sample size* $\sum_{i=1}^{n} \mathbb{1}(X_i Y_i > 0)$, which measures the number of cells with a given perturbation and nonzero expression of a given gene. Table 1 displays the distribution of effective sample sizes across the negative control perturbation-gene pairs tested, showing that the effective samples sizes are vastly smaller than the number of cells, $n = 207{,}324$. Furthermore, Figure 10 stratifies the QQ plots for each method by effective sample size,

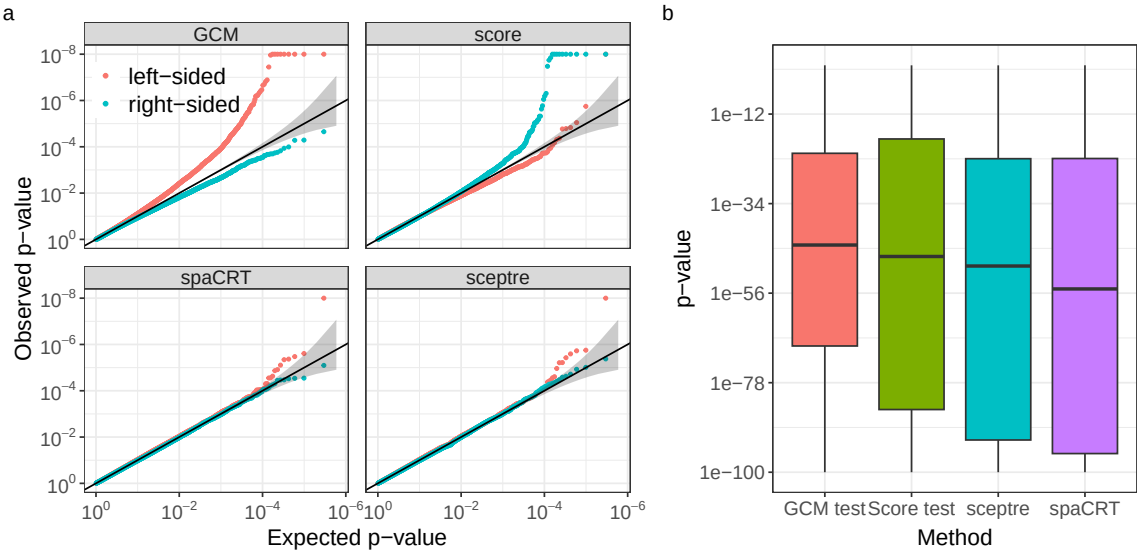

Figure 4: Calibration and power performance in the Gasperini et al. (2019) data. (a): Left- and right-sided p-values for negative control perturbation-gene pairs on the Gasperini data. (b): Left-sided p-values computed on the 754 positive control perturbation-gene pairs in the Gasperini data.

focusing on those pairs with effective sample size of at most 100. As expected based on our simulation study, we find more severe miscalibration for pairs with lower effective sample sizes, especially for the GCM and score tests, and to a lesser extent for `sceptre` and the spaCRT. We carried out a similar analysis, stratifying the pairs based on the estimated size parameter (Figure 11). In line with our simulation results, we find that the GCM and score tests exhibit more miscalibration for smaller size parameters.

	Min.	1st Qu.	Median	3rd Qu.	Max.
Effective sample size	0	53	204	504	2044

Table 1: Effective sample size in subsampled dataset.

Power. Next, we compare the power of the four methods based on their left-sided p-values on the 754 positive control perturbation-gene pairs (Figure 4b). The signal is quite strong in these positive control pairs, as evidenced by small p-values for all four methods. We remark that the spaCRT overcomes the discreteness in the p-values returned by resampling-based methods such as the dCRT, delivering very small p-values in the presence of strong signals. Given the scale of the p-values, we refrain from making definitive conclusions about the relative power of the methods, but remark only that the spaCRT appears at least as powerful as the alternative methods considered.

Computation. The excellent statistical properties of the spaCRT do not come at the cost of computational efficiency. Since the `sceptre` software is highly optimized, while the other methods are not, we benchmark our dCRT implementation (used in

the numerical simulations) instead of `sceptre`'s for computational efficiency. We use $M = 100{,}000$ resamples for the dCRT, given the high multiplicity of the problem (recall Section 1.1). To assess running time, we paired two randomly sampled genes with each of the 51 non-targeting perturbations, for a total of 102 perturbation-gene pairs. We find that the spaCRT is roughly as fast as the GCM test, about five times faster than the score test, and about 250 times faster than the dCRT.

Table 2: Computation time per perturbation-gene pair on the Gasperini data. Times are reported in seconds.

Method	Mean	Std dev
GCM test	4.0	1.9
dCRT	1002.1	279.5
spaCRT	4.0	1.9
Score test	19.9	9.3

7 Discussion

In this paper, we prove a new SPA for approximating resampling distribution and then introduce the spaCRT, a new conditional independence test enjoying several desirable properties building on the new theoretical results. spaCRT is completely resampling-free, applicable to general regression setups and enjoys excellent Type-I error control and power in both numerical simulations and real data analysis. The spaCRT is particularly well-suited to data with low signal-to-noise ratio and large scale, such as single-cell CRISPR screen data analysis.

Despite the attractive properties of the spaCRT, the current work still has several limitations. First, there remains a gap between our theoretical guarantees and the practical performance of the spaCRT and dCRT. We have shown that both of these tests have Type-I error control asymptotically. Except for Theorem 6 where we establish the theoretical gain of odCRT compared to oGCM, we have not theoretically justified why these tests perform so well in finite samples compared with asymptotic tests like the GCM and score tests, when in-sample fitting for $\mu_{n,x}$ and $\mu_{n,y}$ are considered. Second, we have focused on approximating the dCRT in this paper, but the conditional randomization test framework is applicable to general test statistics. It would be interesting to see how the saddlepoint approximation can be extended to other test statistics.

8 Acknowledgments

We acknowledge the Wharton research computing team for their help with our use of the Wharton high-performance computing cluster for the numerical simulations and real data analyses in this paper. This work was partially support by NSF DMS-2113072 and NSF DMS-2310654.

References

Adamson, Britt et al. (2016). "A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response". In: *Cell* 167.7, 1867–1882.e21.

Auer, Paul L and Guillaume Lettre (2015). "Rare variant association studies: considerations, challenges and opportunities". In: *Genome medicine* 7, pp. 1–11.

Barber, Rina Foygel and Emmanuel J Candès (2015). "Controlling the false discovery rate via knockoffs". In: *The Annals of statistics*, pp. 2055–2085.

Barry, Timothy, Kaishu Mason, Kathryn Roeder, and Eugene Katsevich (2024). "Robust differential expression testing for single-cell CRISPR screens at low multiplicity of infection". In: *Genome Biology* 25.1, pp. 1–30.

Barry, Timothy, Xuran Wang, John A. Morris, Kathryn Roeder, and Eugene Katsevich (2021). "SCEPTRE improves calibration and sensitivity in single-cell CRISPR screen analysis". In: *Genome Biology* 22.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1, pp. 289–300.

Bentkus, Vidmantas, Friedrich Götze, and Willem R van Zwet (1997). "An Edgeworth expansion for symmetric statistics". In: *The Annals of Statistics* 25.2, pp. 851–896.

Besag, Julian and Peter Clifford (1991). "Sequential Monte Carlo p-values". In: *Biometrika* 78.2, pp. 301–304.

Bickel, P. J. (1974). "Edgeworth expansions in nonparametric statistics". In: *Annals of Statistics* 2.1, pp. 1–20.

Browning, Sharon R and Brian L Browning (2007). "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering". In: *The American Journal of Human Genetics* 81.5, pp. 1084–1097.

Butler, Ronald (2007). *Saddlepoint approximations with applications*. Cambridge University Press.

Candès, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv (2018). "Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3, pp. 551–577.

Daniels, Henry E. (1954). "Saddlepoint Approximations in Statistics". In: *The Annals of Mathematical Statistics* 25.4, pp. 631–650.

— (1955). "Discussion on the Paper by Dr. Box and Dr. Andersen". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.1, pp. 26–34.

Datlinger, Paul et al. (2017). "Pooled CRISPR screening with single-cell transcriptome readout". In: *Nature Methods* 14.3, pp. 297–301.

Davidson, James (1994). *Stochastic Limit Theory*. Oxford University Press.

Davison, Anthony C . and David V. Hinkley (1988). "Saddlepoint Approximations in Resampling Methods". In: *Biometrika* 75.3, pp. 417–431.

Dey, Rounak, Ellen M Schmidt, Goncalo R Abecasis, and Seunggeun Lee (2017). "A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS". In: *The American Journal of Human Genetics* 101.1, pp. 37–49.

Dixit, Atray et al. (2016). "Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens". In: *Cell* 167, pp. 1853–1866.
Efron, Bradley (2022). *Exponential Families in Theory and Practice*. Cambridge University Press, pp. 1–250.
Fischer, Lasse and Aaditya Ramdas (2024a). "Multiple testing with anytime-valid Monte-Carlo p-values". In: pp. 1–22. arXiv: 2404.15586.
— (2024b). "Sequential Monte-Carlo testing by betting". In: *arXiv*, pp. 1–33. arXiv: 2401.07365.
Gandy, Axel (2009). "Sequential implementation of Monte Carlo tests with uniformly bounded resampling risk". In: *Journal of the American Statistical Association* 104.488, pp. 1504–1511. arXiv: 0612488 [math].
Gandy, Axel and Georg Hahn (2014). "MMCTest-A safe algorithm for implementing multiple monte carlo tests". In: *Scandinavian Journal of Statistics* 41.4, pp. 1083–1101.
— (2016). "A Framework for Monte Carlo based Multiple Testing". In: *Scandinavian Journal of Statistics* 43.4, pp. 1046–1063.
— (2017). "QuickMMCTest: quick multiple Monte Carlo testing". In: *Statistics and Computing* 27.3, pp. 823–832.
Gasperini, Molly et al. (2019). "A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens". In: *Cell* 176.1-2, 377–390.e19.
Ge, Tian, Jianfeng Feng, Derrek P. Hibar, Paul M. Thompson, and Thomas E. Nichols (2012). "Increasing power for voxel-wise genome-wide association studies: The random field theory, least square kernel machines and fast permutation procedures". In: *NeuroImage* 63.2, pp. 858–873.
Hall, Peter (2013). *The bootstrap and Edgeworth expansion*. Springer Science & Business Media.
Hemerik, Jesse and Jelle Goeman (2018). "Exact testing with random permutations". In: *Test* 27.4, pp. 811–825.
Hemerik, Jesse, Jelle J Goeman, and Livio Finos (2020). "Robust testing in generalized linear models by sign-flipping score contributions". In: *Journal of the Royal Statistical Society, Series B* 82.3, pp. 841–864.
Huang, Mo et al. (2018). "SAVER: Gene expression recovery for single-cell RNA sequencing". In: *Nature Methods* 15.7, pp. 539–542.
Imbens, Guido W and Donald B Rubin (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press.
Jaitin, Diego Adhemar et al. (2016). "Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq". In: *Cell* 167.7, 1883–1896.e15.
Kolassa, John E. (2006). *Series Approximation Methods in Statistics*. Third Edit. Springer.
— (2007). "A proof of the asymptotic equivalence of two-tail probability approximations". In: *Communications in Statistics - Theory and Methods* 36.2, pp. 221–228.
Kuchibhotla, Arun Kumar (2023). "Central Limit Theorems and Approximation Theory: Part II". In: *arXiv preprint arXiv:2306.14382*.
Liu, Molei, Eugene Katsevich, Lucas Janson, and Aaditya Ramdas (2022). "Fast and powerful conditional randomization testing via distillation". In: *Biometrika* 109.2, pp. 277–293.

Lugannani, Robert and Stephen Rice (1980). "Saddle point approximation for the distribution of the sum of independent random variables". In: *Advances in Applied Probability* 12.2, pp. 475–490.

Marchini, Jonathan, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly (2007). "A new multipoint method for genome-wide association studies by imputation of genotypes". In: *Nature genetics* 39.7, pp. 906–913.

Niu, Ziang, Abhinav Chakraborty, Oliver Dukes, and Eugene Katsevich (2024). "Reconciling model-X and doubly robust approaches to conditional independence testing". In: *Annals of Statistics, to appear*.

Petrov, Valentin V (Apr. 1995). *Oxford Studies In Probability 4: Limit Theorems of Probability Theory Sequences of Independent Random Variables*. Oxford University Press.

Robinson, J. (1982). "Saddlepoint Approximations for Permutation Tests and Confidence Intervals". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.1, pp. 91–101.

Scheet, Paul and Matthew Stephens (2006). "A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase". In: *The American Journal of Human Genetics* 78.4, pp. 629–644.

Sesia, Matteo, Chiara Sabatti, and Emmanuel J Candès (2019). "Gene hunting with hidden Markov model knockoffs". In: *Biometrika* 106.1, pp. 1–18.

Shah, Rajen D. and Jonas Peters (2020). "The Hardness of Conditional Independence Testing and the Generalised Covariance Measure". In: *Annals of Statistics* 48.3, pp. 1514–1538.

Svensson, Valentine (2020). "Droplet scRNA-seq is not zero-inflated". In: *Nature Biotechnology* 38, pp. 142–150.

Swanson, Jason (2019). *Lecture notes on probability theory*. Tech. rep.

Tibshirani, Robert (1996). "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 58.1, pp. 267–288.

Townes, F. William, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry (2019). "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model". In: *Genome Biology* 20.1, pp. 1–16.

Turro, Ernest et al. (2020). "Whole-genome sequencing of patients with rare diseases in a national health system". In: *Nature* 583.7814, pp. 96–102.

Visscher, Peter M et al. (2017). "10 years of GWAS discovery: biology, function, and translation". In: *The American Journal of Human Genetics* 101.1, pp. 5–22.

Wainwright, Martin J (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press.

Winkler, Anderson M., Gerard R. Ridgway, Gwenaëlle Douaud, Thomas E. Nichols, and Stephen M. Smith (2016). "Faster permutation inference in brain imaging". In: *NeuroImage* 141, pp. 502–516.

Zhao, Zhangchen et al. (2020). "UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test". In: *The American Journal of Human Genetics* 106.1, pp. 3–12.

A Probability theory preliminaries

A.1 Single probability space embedding

To better state and understand the conditional convergence result, the following lemma helps to embed all the random variables into one big probability space.

Lemma 1 (Embedding into a single probability space, Lemma 14 in Niu et al., 2024)**.** *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* $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$*.*

With the above Lemma, we are safe to state any almost sure statement which can be interpreted within one probability space.

A.2 Some facts about natural exponential family

Consider the NEF with probability density function

$$f(x|\theta) = h(x)\exp(\theta x - A(\theta)).$$

Then there is one-to-one correspondence between the moments of the random variable from NEF and the derivative of the log-partition function $A(\theta)$. We summarise the relationship in the following Lemma.

Lemma 2 (Chapter 1.2 in Efron, 2022)**.** *Suppose* $X \sim f(x|\theta)$ *then the following identities hold:*

1. $\mathbb{E}[X] = A'(\theta)$;
2. $\mathbb{E}[X^2] - (\mathbb{E}[X])^2 = A''(\theta)$;
3. $\mathbb{E}[(X - \mathbb{E}[X])^3] = A^{(3)}(\theta)$;
4. $\mathbb{E}[(X - \mathbb{E}[X])^4] - 3\left(\mathbb{E}[(X - \mathbb{E}[X])^2]\right)^2 = A^{(4)}(\theta)$.

B Additional details of Section 2

B.1 Additional details of Section 2.1

Either of the Assumption 1 or Assumption 2 in the previous section guarantees the existence of these CGFs and their derivatives in a neighborhood of the origin.

Lemma 3. *Suppose Assumption 1 or Assumption 2 holds. Then, there exists a probability-one event* $\mathcal{A}$ *and an* $\varepsilon > 0$ *such that, on* $\mathcal{A}$*,*

$$K_{in}(s) < \infty \quad \text{for any } s \in (-\varepsilon, \varepsilon) \text{ and for all } i \leq n,\ n \geq 1 \tag{34}$$

and

$$|K_{in}^{(r)}(s)| < \infty \quad \textit{for any } s \in (-\varepsilon, \varepsilon) \textit{ and for all } i \leq n,\ n \geq 1,\ r \geq 1, r \in \mathbb{N}, \tag{35}$$

where $K_{in}^{(r)}$ *denotes the* r*-th derivative of* K_{in}*.*

We now explicitly define the solution to saddlepoint equation (5) $\hat{s}_n$. In particular, we restrict our attention to solutions in the interval $[-\varepsilon/2, \varepsilon/2]$:

$$S_n \equiv \{s \in [-\varepsilon/2, \varepsilon/2] : K_n'(s) = w_n\}.$$

It is possible, for specific realizations of $K_n'(s)$ and w_n, that the set S_n is either empty or contains multiple elements. To make the saddlepoint approximation well-defined in these cases, we define $\hat{s}_n$ as follows:

$$\hat{s}_n \equiv \begin{cases} \text{the single element of } S_n & \text{if } |S_n| = 1; \\ \frac{\varepsilon}{2}\text{sgn}(w_n) & \text{otherwise.} \end{cases} \tag{36}$$

Note that this definition ensures that $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$.

B.2 Application to sign-flipping test

In this section, we apply Theorem 1 to derive and justify the validity of the Lugannani-Rice SPA for the sign-flipping test. Suppose

$$X_{in} = \mu_n + \varepsilon_{in}, \quad \varepsilon_{in} \overset{\text{ind}}{\sim} F_{in}, \tag{37}$$

where the error distributions F_{in} are symmetric, but potentially distinct and unknown. We are interested in testing

$$H_{0n} : \mu_n = 0 \quad \text{versus} \quad H_{1n} : \mu_n > 0 \tag{38}$$

based on $T_n \equiv \frac{1}{n}\sum_{i=1}^n X_{in}$. Note that the SPA cannot directly be applied to approximate tail probabilities of T_n because the error distributions F_{in} are unknown. Instead, we can approximate the tail probability of T_n by conditioning on the observed data and resampling the signs of the data. In particular, define the resamples

$$\widetilde{X}_{in} \equiv \pi_{in} X_{in}, \quad \pi_{in} \overset{\text{i.i.d.}}{\sim} \text{Rad}(0.5), \tag{39}$$

where Rad(0.5) denotes the Rademacher distribution placing equal probability mass on ± 1. Due to the assumed symmetry of the distributions F_{in}, flipping the signs of X_{in} preserves their distributions under the null hypothesis, guaranteeing finite-sample validity of the resampling-based p-value from equation (3) (Hemerik and Goeman, 2018; Hemerik, Goeman, and Finos, 2020). To circumvent the computationally costly resampling inherent in the sign-flipping test, we can obtain an accurate approximation to the p-value by applying the SPA to the tail probabilities of the resampling distribution

$$\widetilde{T}_n \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{X}_{in} \equiv \frac{1}{n}\sum_{i=1}^n \pi_{in} X_{in}. \tag{40}$$

Such approximations have been proposed before (Daniels, 1955; Robinson, 1982; Davison and Hinkley, 1988), but have not been rigorously justified (see also Section 2.2). We will now apply Theorem 1 to derive and justify the Lugananni-Rice SPA for the sign-flipping test. We derive the saddlepoint approximation $\widehat{\mathbb{P}}_{\text{LR}}$. Defining $\mathcal{F}_n \equiv \sigma(X_{1n}, \ldots, X_{nn})$, we first calculate the conditional cumulant-generating functions

$$K_{in}(s) \equiv \log \mathbb{E}\left[\exp(s\widetilde{X}_{in})|\mathcal{F}_n\right] = \log\left(\frac{\exp(sX_{in}) + \exp(-sX_{in})}{2}\right) = \log\cosh(sX_{in})$$

and their first two derivatives

$$K'_{in}(s) = X_{in} - \frac{2X_{in}}{1+\exp(2sX_{in})} \quad \text{and} \quad K''_{in}(s) = \frac{4X_{in}^2\exp(2sX_{in})}{(1+\exp(2sX_{in}))^2}. \tag{41}$$

Therefore, the saddlepoint equation (5) reduces to

$$\frac{1}{n}\sum_{i=1}^n K'_{in}(s) = \frac{1}{n}\sum_{i=1}^n X_i \quad \Longleftrightarrow \quad \sum_{i=1}^n \frac{X_{in}}{1+\exp(2sX_{in})} = 0. \tag{42}$$

Given a solution $\hat{s}_n$ to the saddlepoint equation (whose existence and uniqueness is guaranteed by Theorem 5 below), we can define the quantities λ_n and r_n from equation (6):

$$\lambda_n \equiv \hat{s}_n\sqrt{nK''_n(\hat{s}_n)} = \hat{s}_n\sqrt{\sum_{i=1}^n \frac{4X_{in}^2\exp(2\hat{s}_nX_{in})}{(1+\exp(2\hat{s}_nX_{in}))^2}} \tag{43}$$

and

$$r_n \equiv \operatorname{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_nw_n - K_n(\hat{s}_n))} = \operatorname{sgn}(\hat{s}_n)\sqrt{2\sum_{i=1}^n(\hat{s}_nX_{in} - \log\cosh(\hat{s}_nX_{in}))}, \tag{44}$$

where we set $r_n \equiv \operatorname{sgn}(\hat{s}_n)$ when the quantity under the square root is negative. With these definitions, the SPA for the tail probability of interest is

$$\widehat{\mathbb{P}}_{\text{LR}}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^n X_{in} \;\middle|\; \mathcal{F}_n\right] \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \tag{45}$$

The following theorem gives sufficient conditions for this saddlepoint approximation to have vanishing relative error.

Theorem 5. *Suppose X_{in} are drawn from the probability model (37), such that*

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_{in}^2] > 0; \tag{46}$$

$$\text{there exists } \delta > 0 \text{ such that } \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \mathbb{E}[|\varepsilon_{in}|^{4+\delta}] < \infty; \tag{47}$$

$$\mu_n = o(1). \tag{48}$$

Then, the saddlepoint equation (42) *has a unique solution* $\hat{s}_n \in [-1, 1]$ *with probability approaching 1 as* $n \to \infty$. *Furthermore, the tail probability approximation* (45) *obtained from equations* (43) *and* (44) *has vanishing relative error:*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^{n}X_{in} \mid \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^{n}X_{in} \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \quad (49)$$

C Additional details of Section 3.1

C.1 Doubly robust properties of dCRT and GCM test

dCRT is a resampling-based procedure that can be computationally challenging when the sample size is moderate to large. An alternative test procedure is instead of using the resampling to construct p-value, one can use the normal approximation to the test statistic $T_n^{\mathrm{dCRT}}(X, Y, Z)$ (17) under null and construct a p-value based on the cutoff of the standard normal distribution after properly normalizing the standard deviation estimate. This is the so-called *generalized covariance measure* (GCM) test (Shah and Peters, 2020). Building on the same test statistic (except for extra normalization), GCM also allows flexible modeling choices on estimators $\widehat{\mu}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,y}(\cdot)$. In fact, it has been proved in Shah and Peters (2020) that GCM enjoys the so-called double robustness property: as long as both estimators $\widehat{\mu}_{n,x}(\cdot)$ and $\widehat{\mu}_{n,y}(\cdot)$ are consistent and converge to the true conditional expectations at rate faster than $n^{-1/4}$, the validity of the test can be guaranteed (Shah and Peters, 2020, Theorem 6).

Proved in Niu et al. (2024), the dCRT is asymptotically equivalent to GCM under mild conditions. Thus dCRT inherits the doubly robust statistical property from GCM. One important difference is dCRT relies on resampling to construct the p-value, while GCM relies on the asymptotic normal approximation. This difference can lead to advantage for GCM for analyzing data with large scale because of fast computation.

C.2 Illustrative results with oracle knowledge of $\mu_{n,x}$ and $\mu_{n,y}$

To develop the intuition, we work with a simple but illustrative setup where we assume that the conditional expectations $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$ are known and we will set $\widehat{\mu}_{n,x}(\cdot) = \mu_{n,x}(\cdot)$ and $\widehat{\mu}_{n,y}(\cdot) = \mu_{n,y}(\cdot)$ in the test statistics. We will focus on a Bernoulli model for $\boldsymbol{X}|\boldsymbol{Z}$:

$$\boldsymbol{X}|\boldsymbol{Z} \sim \mathrm{Ber}(\mu_{n,x}(\boldsymbol{Z})). \quad (50)$$

We define the *oracle* GCM (oGCM) test by considering the test statistic

$$\phi_{n,\alpha}^{\mathrm{oGCM}} \equiv \mathbb{1}(T_n^{\mathrm{oGCM}}(X, Y, Z) > z_{1-\alpha}) \quad \text{where} \quad T_n^{\mathrm{oGCM}}(X, Y, Z) \equiv \frac{1}{nS_n}\sum_{i=1}^{n} R_{in}^o, \quad (51)$$

$R_{in}^o \equiv (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))$ and $S_n^2 = \mathbb{E}[(R_{in}^o)^2]$. For dCRT, we consider the modified dCRT with theoretical quantile, which we call oracle dCRT (odCRT):

$$\phi_{n,\alpha}^{\mathrm{odCRT}} \equiv \mathbb{1}\left(T_n^{\mathrm{dCRT}} \geq \mathbb{Q}_{1-\alpha}(\tilde{T}_n^{\mathrm{dCRT}}|X, Y, Z)\right)$$

where $T_n^{\mathrm{dCRT}}, \tilde{T}_n^{\mathrm{dCRT}}$ are defined in (17) and (18) respectively, but with oracle $\mu_{n,x}(\cdot), \mu_{n,y}(\cdot)$ and resampling distribution $\widetilde{X}^{(m)} \sim \prod_{i=1}^n \mathrm{Ber}(\mu_{n,x}(Z_{in}))$. The intuition for the Type-I error deviating from the specified significance lelve for GCM (and oGCM) is the CLT, when excessive sparsity exists, can happen in an arbitrarily slow rate depending how sparse the data is. To formalize such intuition, we consider the following assumptions.

Assumption 6 (Sparsity level in $\boldsymbol{X}$)**.** *Suppose v_n is a sequence of positive constants and $cv_n \leq \inf_z |\mu_{n,x}(z)| \leq \sup_z |\mu_{n,x}(z)| \leq Cv_n$ for some universal constants $C > 0$ and $c > 0$.*

Assumption 7 (Conditional moments of $\boldsymbol{Y}$)**.** *Suppose the following conditions hold:* $\sup_z \mathbb{E}[\boldsymbol{Y}^4|\boldsymbol{Z} = z] < \infty$, $\inf_z \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}])^3|\boldsymbol{Z} = z] > 0$ *and* $\inf_z \mathrm{Var}[\boldsymbol{Y}|\boldsymbol{Z} = z] > 0$.

Assumption 8 (Cramér's condition)**.** *Suppose $S_{in} \equiv R_{in}^o / \sqrt{\mathbb{E}[(S_n)^2]}$ satisfies the Cramér's condition:* $\limsup_{n\to\infty} \sup_{t\in\mathbb{R}} |\mathbb{E}[\exp(itS_{in})]| < 1$.

Parameter v_n in Assumption 6 characterizes the sparsity level of data, which will play an important role in Theorem 6 to unveil the failure of oGCM to control Type-I error under small sample size. Assumption 7 states the bounded moment condition for $\boldsymbol{Y}$ given $\boldsymbol{Z}$ as well as the non-degeneracy of the conditional variance and conditional third central moment. This is mainly required to prove the rate of convergece on Type-I error for oGCM test. Such assumption can be staisfied by examples including Poisson or negative binomial case with uniformly lower and upper bounded conditional mean (and fixed dispersion parameter for negative binomial case) in $\boldsymbol{Y}|\boldsymbol{Z}$. This corresponds to the setup in Figure 1. Assumption 8 is used to guarantee the validity of *Edgeworth expansion* on S_{in}. The assumption may seem to be contradictory with model setup (50) at the first glance because of potential sparsity in X. However, this is not the case. The key reason hinges on the convolution nature of random variable S_{in} and as long as $\mu_{n,x}(Z_{in}) \cdot \mu_{n,y}(Z_{in})$ are continuous random variables, the convolution of the product variable with discrete random variables can still satisfy the Cramér's condition. Now we state our illustrative results.

Theorem 6. *Consider $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$. Suppose Assumption 6-8 hold. Then we have*

1. ***Finite-sample validity of*** odCRT***:*** $\mathbb{E}[\phi_{n,\alpha}^{\mathrm{odCRT}}] = \alpha$*;*
2. ***Convergence of Type-I error of*** oGCM***:*** *If $1/v_n = o(n)$, then there exists a sequence $r_n > 0$ such that $r_n \sim 1/(nv_n)^{1/2}$ and*

$$\left|\mathbb{E}[\phi_{n,\alpha}^{\mathrm{oGCM}}] - \alpha - r_n\right| = o(r_n). \tag{52}$$

The argument for finite-sample validity is by the exchangeability of the resampled data and the original data under null hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$. Now we discuss the implication of the results on oGCM.

Remark 7 (Implication for testing with oGCM)**.** Theorem 6 unveils that sparsity in data can slow down the Type-I error converging to the specified significance level α, which is a result of slow convergence rate of CDF of oGCM test statistic converging to the CDF of standard normal distribution. When v_n is of order n^{-s} for $s > 0$, the convergence rate of Type-I error of oGCM is $n^{(1-s)/2}$. The closer s is to 1, the slower the convergence rate is.

Theorem 6 considers the model-X assumption for odCRT and oracle knowledge of $\mu_{n,x}, \mu_{n,y}$. Thus it only serves as a illustration and the results are not directly applicable to the general case where $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$ are unknown. However, the theorem provides a high-level insight on the finite-sample performance of dCRT and GCM tests.

C.3 Proof of Theorem 6

To prove Theorem 6, we first need an auxiliary result.

Lemma 4 (Theorem 5.18 in (Petrov, 1995); Theorem 4.1 in (Kuchibhotla, 2023))**.** *Consider a sequence of independently and identically distributed random variables $W_{in} \in \mathbb{R}$. Suppose $\mathbb{E}[W_{in}] = 0, \mathbb{E}[W_{in}^2] = 1$ and $\mathbb{E}[W_{in}^4] < \infty$ for any $n \in \mathbb{N}$. Then there exists a universal constant $C > 0$ such that for all $x \in \mathbb{R}$,*

$$\left| \mathbb{P}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} \leq x \right] - \mathbb{P}[Z \leq x] - \frac{(1-x^2)\exp(-x^2/2)\mathbb{E}[W_{in}^3]}{6\sqrt{2\pi n}} \right| \leq C\frac{\mathbb{E}[W_{in}^4]}{n} + C\left(\sup_{|t| \geq 1/(12\mathbb{E}[|W_{in}|^3])} |\mathbb{E}[\exp(itW_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1+|x|^4}. \tag{53}$$

Proof of Theorem 6. This argument for the validity of odCRT is based on the exchangeability of the resampled data and the original data under null hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$. Now we prove the convergence rate of oGCM. In order to prove the result, we will use Lemma 4 to state an asymptotic expansion of CDF of oGCM test statistic (51). We apply $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ in Lemma 4 so that we get a bound as in (52) with explicit r_n defined as

$$r_n = \frac{(1-z_{1-\alpha}^2)\exp(-z_{1-\alpha}^2/2)\mathbb{E}[S_{in}^3]}{6\sqrt{2\pi n}}.$$

We just need to show that (53) with $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ is of smaller order of r_n. It is sufficient to show the following results:

$$\frac{\mathbb{E}[S_{in}^4]}{nr_n} = o(1) \quad \text{and} \quad \frac{1}{r_n}\left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1+|z_{1-\alpha}|^4} = o(1).$$

To prove these statements, we first show the order of convergence of r_n. Then we can obtain by conditional independence $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$ and Assumption 7 that

$$\begin{aligned} r_n &\sim \frac{1}{n^{1/2}} \frac{\mathbb{E}[\mu_{n,x}(Z_{in})(1-\mu_{n,x}(Z_{in}))(1-2\mu_{n,x}(Z_{in}))\mathbb{E}[(Y_{in}-\mathbb{E}[Y_{in}|Z_{in}])^3|Z_{in}]]}{(\mathbb{E}[\mu_{n,x}(Z_{in})(1-\mu_{n,x}(Z_{in}))\mathrm{Var}[Y_{in}|Z_{in}]])^{3/2}} \\ &\sim \frac{1}{n^{1/2}v_n^{1/2}}. \end{aligned}$$

The last equivalence holds by Hölder's inequality. Together with Assumption 8, we know

$$\limsup_{n\to\infty} \sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| \leq \limsup_{n\to\infty} \sup_{t\in\mathbb{R}} |\mathbb{E}[\exp(itS_{in})]| \equiv c < 1.$$

Thus we have for any $\gamma > 0$,

$$\left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \lesssim (c + 1/2n)^n = o(n^\gamma)$$

so that we prove

$$\frac{1}{r_n} \left(\sup_{|t| \geq 1/(12\mathbb{E}[|S_{in}|^3])} |\mathbb{E}[\exp(itS_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1 + |z_{1-\alpha}|^4} = o(1).$$

Then it remains to prove

$$\frac{\mathbb{E}[S_{in}^4]}{n r_n} = o(1). \tag{54}$$

Proof of convergence (54): It suffices to show $\mathbb{E}[S_{in}^4] = o(n^{1/2}/v_n^{1/2})$ by the proved results $r_n \sim 1/(nv_n)^{1/2}$. To see this, by Assumption 7, we can bound

$$\begin{aligned}
\mathbb{E}[S_{in}^4] &\leq \frac{\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))(1 - 3\mu_{n,x}(Z_{in}) + 3\mu_{n,x}^2(Z_{in}))]}{(\mathbb{E}[\mu_{n,x}(Z_{in})(1 - \mu_{n,x}(Z_{in}))])^2} \\
&\quad \times \frac{\sup_z \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^4 | Z_{in} = z]}{\inf_z \mathrm{Var}^2[Y_{in}|Z_{in} = z]} \\
&\sim \frac{1}{v_n}.
\end{aligned}$$

Then by the assumption that $1/v_n = o(n)$, we know $\mathbb{E}[S_{in}^4] = o(n^{1/2}/v_n^{1/2})$. □

D Some useful lemmas and proofs

Lemma 5 (Lemma 3 in Niu et al., 2024)**.** *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:

$$C_n(X, Y, Z) \xrightarrow{\mathbb{P}} z_{1-\alpha}$$

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

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - z_{1-\alpha}| \leq \delta] = 0, \tag{55}$$

then the two tests are asymptotically equivalent:

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

Regularity condition: there exists $\delta > 0$ such that for a sequence of laws $\mathcal{L}_n$ and its estimate $\widehat{\mathcal{L}}_n$, the following assumptions hold:

$$(\widehat{S}_n^{\mathrm{dCRT}})^2 \equiv \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in} \mid Z_{in}](Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 = \Omega_{\mathbb{P}}(1); \tag{56}$$

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n} |Y_{in} - \widehat{\mu}_{n,y}(Z_{in})|^{2+\delta}\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^{2+\delta} \mid X, Z] = o_{\mathbb{P}}(1); \tag{57}$$

$$\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in}|Z_{in}], (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2, (Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty \text{ almost surely.} \tag{58}$$

Lemma 6 (Theorem 9 in Niu et al., 2024)**.** *Let $\mathcal{L}_n$ be a sequence of laws and $\widehat{\mathcal{L}}_n$ be a sequence of estimates. Suppose there exists a sequence of laws $\mathcal{L}_n$ satisfying all the assumptions in* ***Regularity condition****. Then, the quantile of*

$$T_n^{\mathrm{ndCRT}}(\widetilde{X}, X, Y, Z) \equiv \frac{T_n^{\mathrm{dCRT}}(\widetilde{X}, X, Y, Z)}{\widehat{S}_n^{\mathrm{dCRT}}} \tag{59}$$

converges to the quantile of the standard normal distribution pointwisely in probability, i.e., for any $p \in (0, 1)$,

$$\mathbb{Q}_p\left[n^{1/2}T_n^{\mathrm{ndCRT}}(\widetilde{X}, X, Y, Z)|X, Y, Z\right] \xrightarrow{\mathbb{P}} z_p.$$

Lemma 7 (Corollary 6 in Niu et al., 2024)**.** *Let X_{in} be a triangular array of random variables, such that X_{in} are independent for each n. If for some $\delta > 0$ we have*

$$\frac{1}{n^{1+\delta}}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^{1+\delta}] \to 0, \tag{60}$$

then

$$\frac{1}{n}\sum_{i=1}^{n}(X_{in} - \mathbb{E}[X_{in}]) \xrightarrow{\mathbb{P}} 0.$$

The condition (60) is satisfied when

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

Lemma 8 (Dominance of higher moment)**.** *For any $1 < p < q < \infty$, the following inequality is true almost surely:*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^p|\mathcal{F}_n] \le \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^q|\mathcal{F}_n]\right)^{p/q}.$$

Lemma 9 (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*

$$\mathbb{E}[|W_1W_2| \mid \mathcal{F}] \le (\mathbb{E}[|W_1|^{q_1} \mid \mathcal{F}])^{1/q_1}(\mathbb{E}[|W_2|^{q_2} \mid \mathcal{F}])^{1/q_2} \quad \textit{almost surely.}$$

Lemma 10 (Conditional Jensen inequality, Davidson, 1994, 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*

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

Lemma 11 (Conditional Markov's inequality, Davidson, 1994, 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*

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

Lemma 12 (Conditional central limit theorem)**.** *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*

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

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

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

then

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

Lemma 13. *Consider the fixed dimension setup: $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$. Define $\sigma^2_{\mathrm{dCRT}} = \mathbb{E}[(X_i - \mathbb{E}[X_i|Z_i])^2 (Y_i - \mathbb{E}[Y_i|Z_i])^2]$. Then if $\mathbb{E}[(Y_i - \mathbb{E}[Y_i|Z_i])^2] < \infty$ and $\mathbb{P}[X_i, \widehat{\mu}_x(Z_i) \in [-S, S]] = 1$ for some $S > 0$, then as long as the following conditions hold:*

$$\begin{aligned}
\boldsymbol{X} &\perp\!\!\!\perp \boldsymbol{Y} | \boldsymbol{Z}; \\
\sigma^2_{\mathrm{dCRT}} &\in (0, \infty); \\
\frac{1}{n} \sum_{i=1}^{n} (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 &= o_{\mathbb{P}}(1); \\
\frac{1}{n} \sum_{i=1}^{n} (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 &= o_{\mathbb{P}}(1); \\
\left(\frac{1}{n} \sum_{i=1}^{n} (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 \right) \left(\frac{1}{n} \sum_{i=1}^{n} (\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 \right) &= o_{\mathbb{P}}(1/n),
\end{aligned}$$

then we have $\sqrt{n} T_n^{\mathrm{dCRT}} \xrightarrow{\mathbb{P}} N(0, \sigma^2_{\mathrm{dCRT}})$.

D.1 Proof of Lemma 8

Proof of Lemma 8. By Lemma 9, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^p|\mathcal{F}_n] \le \frac{1}{n}\left(\sum_{i=1}^{n}(\mathbb{E}[|X_{in}|^p|\mathcal{F}_n])^{q/p}\right)^{p/q} n^{1-p/q}$$
$$= \left(\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[|X_{in}|^p|\mathcal{F}_n])^{q/p}\right)^{p/q}.$$

We use Jensen's inequality, Lemma 10, to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^p|\mathcal{F}_n] \le \left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^q|\mathcal{F}_n]\right)^{p/q}.$$

□

D.2 Proof of Lemma 13

We consider the following decomposition:

$$\begin{aligned}T_n^{\mathrm{dCRT}} &= \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_x(Z_i))(Y_i-\mu_y(Z_i)) + \frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))(Y_i-\mu_y(Z_i))\\ &+ \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_x(Z_i))(\mu_y(Z_i)-\widehat{\mu}_y(Z_i)) + \frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))\\ &\equiv \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_x(Z_i))(Y_i-\mu_y(Z_i)) + \sum_{k=1}^{3}B_k.\end{aligned}$$

We will prove $\sqrt{n}B_k$ converges 0 in probability for any $k = 1, 2, 3$. For B_1 and B_2, we have

$$\mathbb{P}\left[\sqrt{n}B_1 > \varepsilon|X, Z\right] \le \frac{1}{\varepsilon^2 n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2\mathbb{E}[(Y_i-\mu_y(Z_i))^2|Z] = o_{\mathbb{P}}(1)$$

and

$$\mathbb{P}\left[\sqrt{n}B_2 > \varepsilon|X, Z\right] \le \frac{1}{\varepsilon^2 n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^2\mathbb{E}[(Y_i-\mu_x(Z_i))^2|Z] = o_{\mathbb{P}}(1).$$

As for B_3, we know by Cauchy-Schwarz inequality that

$$\sqrt{n}|B_3| \le \sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^2} = o_{\mathbb{P}}(1).$$

Therefore, we only need to prove the weak convergence of $\sum_{i=1}^{n}(X_i-\mu_x(Z_i))(Y_i-\mu_y(Z_i))/\sqrt{n}$. This is true by usual CLT.

E Asymptotic equivalence and proofs of results in section 4

E.1 Statement of an asymptotic equivalence result

We first state a result on the asymptotic equivalence of spaCRT and dCRT. This is a generalization of Corollary 2.

Theorem 7. *Suppose the assumptions of Theorem 2 hold. Fix $\alpha \in (0,1)$. If the normalized test statistic $n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)/\widehat{S}_n^{\mathrm{dCRT}}$, where $\widehat{S}_n^{\mathrm{dCRT}}$ is defined in equation (56) (in Appendix), does not accumulate around the $1-\alpha$ quantile of standard normal distribution $z_{1-\alpha}$, i.e.,*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}\left[\left|\frac{n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)}{\widehat{S}_n^{\mathrm{dCRT}}}-z_{1-\alpha}\right|\le\delta\right]=0, \tag{64}$$

then $\lim_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}\left[\phi_{n,\alpha}^{\mathrm{spaCRT}}=\phi_{n,\alpha}^{\mathrm{dCRT}}\right]=1$.

We will restate a stronger version of Theorem 7.

Theorem 8 (Stronger version of Theorem 7)**.** *Suppose the assumptions of Theorem 7 hold. Then define the asymptotic test:*

$$\phi_{n,\alpha}^{\mathrm{asy}}\equiv\mathbb{1}\left(\frac{n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)}{\widehat{S}_n^{\mathrm{dCRT}}}>z_{1-\alpha}\right). \tag{65}$$

Then we have

$$\lim_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}\left[\phi_{n,\alpha}^{\mathrm{spaCRT}}=\phi_{n,\alpha}^{\mathrm{dCRT}}=\phi_{n,\alpha}^{\mathrm{asy}}\right]=1.$$

Consequently, if $n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)/\widehat{S}_n^{\mathrm{dCRT}}\xrightarrow{d}N(0,1)$, we have $\lim_{n\to\infty}\mathbb{E}[\phi_{n,\alpha}^{\mathrm{spaCRT}}]=\alpha$.

E.2 Proof of Theorem 2

We have conditional CGF

$$K_{in}(s|\mathcal{F}_n)=A(\widehat{\theta}_{n,x}(Z_{in})+a_{in}s)-A(\widehat{\theta}_{n,x}(Z_{in}))-a_{in}sA'(\widehat{\theta}_{n,x}(Z_{in})). \tag{66}$$

Then we can compute the CCGF under model (16) as

$$K_n(s\mid\mathcal{F}_n)=\frac{1}{n}\sum_{i=1}^{n}\left\{A(\widehat{\theta}_{n,x}(Z_{in})+a_{in}s)-A(\widehat{\theta}_{n,x}(Z_{in}))-a_{in}sA'(\widehat{\theta}_{n,x}(Z_{in}))\right\}. \tag{67}$$

The first two derivatives of this quantity are

$$K_n'(s\mid\mathcal{F}_n)=\frac{1}{n}\sum_{i=1}^{n}a_{in}\left(A'(\widehat{\theta}_{n,x}(Z_{in})+a_{in}s)-A'(\widehat{\theta}_{n,x}(Z_{in}))\right), \tag{68}$$

$$K_n''(s\mid\mathcal{F}_n)=\frac{1}{n}\sum_{i=1}^{n}a_{in}^2A''(\widehat{\theta}_{n,x}(Z_{in})+a_{in}s). \tag{69}$$

We will apply Theorem 1 and thus verify the conditions in the theorem. We first verify the variance condition (8).

Verification of (8): Compute $K''_{in}(s \mid \mathcal{F}_n) = a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s)$. Then it suffices to guarantee

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[W_{in}^2|\mathcal{F}_n] = \frac{1}{n}\sum_{i=1}^{n} K''_{in}(0 \mid \mathcal{F}_n) = \frac{1}{n}\sum_{i=1}^{n} a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in})) = \Omega_{\mathbb{P}}(1).$$

Next we verify assumption 1 and condition (CCS) respectively.

Verification of Assumption 1 with condition (CSE) and (22): We denote the conditional upper tail probability and lower probability respectively as

$$L_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}\left[X \leq a|\mathcal{F}_n\right], \; U_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}\left[X \geq a|\mathcal{F}_n\right].$$

By condition (CSE), we can compute

$$\begin{aligned}
&\mathbb{P}[W_{in} \geq t|\mathcal{F}_n] \\
&= \mathbb{P}[a_{in}(\widetilde{X}_{in} - A'(\widehat{\theta}_{n,x}(Z_{in}))) \geq t|\mathcal{F}_n] \\
&= \mathbb{1}(a_{in} > 0) U_{\widetilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\widehat{\theta}_{n,x}(Z_{in}))\right) + \mathbb{1}(a_{in} < 0) L_{\widetilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\widehat{\theta}_{n,x}(Z_{in}))\right)
\end{aligned}$$

Then by the definition of natural exponential family, we can write

$$\begin{aligned}
&\mathbb{1}(a_{in} > 0) U_{\widetilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}} + A'(\widehat{\theta}_{n,x}(Z_{in}))\right) \\
&= \mathbb{1}(a_{in} > 0) \int_{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp(\widehat{\theta}_{n,x}(Z_{in})x - A(\widehat{\theta}_{n,x}(Z_{in})))h(x)dx \\
&= \mathbb{1}(a_{in} > 0) \int_{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp(\widehat{\theta}_{n,x}(Z_{in})x + a_{in}x - a_{in}x - A(\widehat{\theta}_{n,x}(Z_{in})))h(x)dx \\
&\leq \mathbb{1}(a_{in} > 0) \int_{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp((\widehat{\theta}_{n,x}(Z_{in}) + a_{in})x - A(\widehat{\theta}_{n,x}(Z_{in})))h(x)\mathrm{d}x \\
&\qquad \times \exp(-t - a_{in}A'(\widehat{\theta}_{n,x}(Z_{in}))) \\
&\leq \mathbb{1}(a_{in} > 0) \int_{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}^{\infty} \exp((\widehat{\theta}_{n,x}(Z_{in}) + a_{in})x - A(a_{in} + \widehat{\theta}_{n,x}(Z_{in})))h(x)\mathrm{d}x \\
&\qquad \times \exp(A(a_{in} + \widehat{\theta}_{n,x}(Z_{in})) - A(\widehat{\theta}_{n,x}(Z_{in}))) \exp(-t - a_{in}A'(\widehat{\theta}_{n,x}(Z_{in}))) \\
&\leq \mathbb{1}(a_{in} > 0) \exp(A(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}) - A(\widehat{\theta}_{n,x}(Z_{in})) - a_{in}A'(\widehat{\theta}_{n,x}(Z_{in}))) \exp(-t) \\
&\leq \mathbb{1}(a_{in} > 0) \exp(|A(\widehat{\theta}_{n,x}(Z_{in}) + a_{in})| + |A(\widehat{\theta}_{n,x}(Z_{in}))| + |a_{in}||A'(\widehat{\theta}_{n,x}(Z_{in}))|) \exp(-t)
\end{aligned}$$

Similarly, we can derive the upper bound for the lower tail Probability:

$$
\begin{aligned}
&\mathbb{1}(a_{in}<0)L_{\widetilde{X}_{in},\mathcal{F}_n}\left(\frac{t}{a_{in}}+A'(\widehat{\theta}_{n,x}(Z_{in}))\right)\\
&=\mathbb{1}(a_{in}<0)\int_{-\infty}^{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}\exp(\widehat{\theta}_{n,x}(Z_{in})x-A(\widehat{\theta}_{n,x}(Z_{in})))h(x)dx\\
&=\mathbb{1}(a_{in}<0)\int_{-\infty}^{t/a_{in}+A'(\widehat{\theta}_{n,x}(Z_{in}))}\exp(\widehat{\theta}_{n,x}(Z_{in})x+a_{in}x-a_{in}x-A(\widehat{\theta}_{n,x}(Z_{in})))h(x)dx\\
&\le\mathbb{1}(a_{in}<0)\exp(A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})-A(\widehat{\theta}_{n,x}(Z_{in}))-a_{in}A'(\widehat{\theta}_{n,x}(Z_{in})))\exp(-t)\\
&\le\mathbb{1}(a_{in}<0)\exp(|A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})|+|A(\widehat{\theta}_{n,x}(Z_{in}))|+|a_{in}||A'(\widehat{\theta}_{n,x}(Z_{in}))|)\exp(-t).
\end{aligned}
$$

Then we have for any $t>0$,

$$
\begin{aligned}
&\mathbb{P}[W_{in}\ge t|\mathcal{F}_n]\\
&\le\exp(|A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})|+|A(\widehat{\theta}_{n,x}(Z_{in}))|+|a_{in}||A'(\widehat{\theta}_{n,x}(Z_{in}))|)\exp(-t)\\
&\le\exp(\sup_i|A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})|+\sup_i|A(\widehat{\theta}_{n,x}(Z_{in}))|+\sup_i|a_{in}||A'(\widehat{\theta}_{n,x}(Z_{in}))|)\exp(-t).
\end{aligned}
$$

Choosing

$$
\theta_n=\exp\left(\sup_i|A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})|+\sup_i|A(\widehat{\theta}_{n,x}(Z_{in}))|+\sup_i|a_{in}||A'(\widehat{\theta}_{n,x}(Z_{in}))|\right)
$$

and $\beta=1$, we need to verify

$$
\theta_n=O_{\mathbb{P}}(1)\text{ and }\theta_n<\infty,\text{ almost surely.}
$$

Since by condition (22), we know $\sup_i|a_{in}|\le\sup_i|Y_{in}|+\sup_i|\widehat{\mu}_{n,y}(Z_{in})|<\infty$ almost surely and $|\widehat{\theta}_{n,x}(Z_{in})|<\infty$ almost surely, we know $\theta_n<\infty$ almost surely. Now we prove $\theta_n=O_{\mathbb{P}}(1)$. By condition (CSE), we know for any fixed $\delta>0$, there exists $M(\delta)>0$ such that

$$
\mathbb{P}\left[\mathcal{S}\right]\ge1-\delta,\text{ where }\mathcal{S}\equiv\left\{\sup_i|\widehat{\theta}_{n,x}(Z_{in})|,\ \sup_i|a_{in}|\in[0,M(\delta)]\right\}.
$$

Then on the event $\mathcal{S}$, we know

$$
\sup_i|A(\widehat{\theta}_{n,x}(Z_{in})+a_{in})|\le\sup_{x\in[-2M(\delta),2M(\delta)]}|A(x)|
$$

and

$$
\sup_i|A'(\widehat{\theta}_{n,x}(Z_{in}))|\le\sup_{x\in[-M(\delta),M(\delta)]}|A'(x)|.
$$

Similarly, on the event $\mathcal{S}$, we have

$$
\sup_i|a_{in}|\le M(\delta),\ \sup_i|A(\widehat{\theta}_{n,x}(Z_{in}))|\le\sup_{x\in[-2M(\delta),2M(\delta)]}|A(x)|.
$$

Therefore we have

$$
\mathbb{P}\left[\theta_n\le\exp\left(2\sup_{x\in[-2M(\delta),2M(\delta)]}|A(x)|+M(\delta)\sup_{x\in[-M(\delta),M(\delta)]}|A'(x)|\right)\right]\ge\mathbb{P}[\mathcal{S}]\ge1-\delta.
$$

Therefore we have $\theta_n=O_{\mathbb{P}}(1)$. Thus $\varepsilon=\beta/8=1/8$ according to Lemma 3.

Verification of Assumption 2 with condition (22) and (CCS): By condition (CCS), we know

$$\mathbb{1}(\widetilde{X}_{in} \in [-S, S]) = 1, \text{ almost surely.}$$

This implies that for any $F \in \mathcal{F}_n$, we have

$$\int_F \mathbb{1}\left(\widetilde{X}_{in} \in [-S, S]\right) d\mathbb{P} = \int_F 1 \mathrm{d}\mathbb{P}.$$

Thus we know

$$\mathbb{P}[\widetilde{X}_{in} \in [-S, S]|\mathcal{F}_n] = \mathbb{E}[\mathbb{1}(\widetilde{X}_{in} \in [-S, S])|\mathcal{F}_n] = 1, \text{ almost surely.}$$

Thus we have

$$\mathbb{P}\left[a_{in}(\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})) \in [-2|a_{in}|S, 2|a_{in}|S]|\mathcal{F}_n\right] = 1, \text{ almost surely.}$$

Then again by condition (22), we know

$$|a_{in}| \leq |Y_{in}| + |\widehat{\mu}_{n,x}(Z_{in})| < \infty, \text{ almost surely.}$$

Moreover, by condition (CCS), we know

$$\frac{1}{n}\sum_{i=1}^{n} 16S^4 a_{in}^4 = \frac{16S^4}{n}\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1).$$

Choosing $\nu_{in} = 2|a_{in}|S$, we complete the proof for CCS distribution. Thus $\varepsilon = 1/8$ according to Lemma 3.

E.3 Proof of Theorem 7 and Theorem 8

Proof of Theorem 7 and Theorem 8. Define the auxiliary test

$$\phi_{n,\alpha}^{\text{aux}} \equiv \mathbb{1}\left(\frac{n^{1/2}T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}} > \mathbb{Q}_{1-\alpha(1+M_n)}\left[n^{1/2}T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z)|X, Y, Z\right]\right)$$

where $T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z)$ is defined in (59) and M_n is the sequence such that

$$\mathbb{P}\left(T_n^{\text{dCRT}}(\widetilde{X}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z)|\mathcal{F}_n\right) = p_{\text{spaCRT}}(1 + M_n), \text{ almost surely.}$$

Notice the tests $\phi_{n,\alpha}^{\text{spaCRT}}$ and $\phi_{n,\alpha}^{\text{aux}}$ are equivalent as long as $M_n \in (-1, 1/\alpha - 1)$. Indeed, we have

$$\mathbb{P}[\phi_{n,\alpha}^{\text{spaCRT}} \neq \phi_{n,\alpha}^{\text{aux}}] \leq \mathbb{P}[\widehat{S}_n^{\text{dCRT}} = 0] + \mathbb{P}[M_n \notin (-1, 1/\alpha - 1)] \to 0$$

due to assumption (56) and that $M_n = o_{\mathbb{P}}(1)$. Thus we only need to verify $\phi_{n,\alpha}^{\text{aux}}$ and $\phi_{n,\alpha}^{\text{dCRT}}$ are asymptotically equivalent. Recall the asymptotic test (65). We will prove $\phi_{n,\alpha}^{\text{aux}}, \phi_{n,\alpha}^{\text{asy}}$ are asymptotically equivalent and $\phi_{n,\alpha}^{\text{dCRT}}, \phi_{n,\alpha}^{\text{asy}}$ are asymptotically equivalent, respectively.

Proof of the equivalence of $\phi^{\text{aux}}_{n,\alpha}, \phi^{\text{asy}}_{n,\alpha}$: We apply Lemma 5 with the test statistic $T_n(X, Y, Z)$ to be

$$T_n(X, Y, Z) \equiv \frac{n^{1/2} T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\text{dCRT}}}$$

and cutoff $C_n(X, Y, Z)$ to be

$$C_n(X, Y, Z) \equiv \mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z) | X, Y, Z \right].$$

The following lemma characterizes the convergence of $C_n(X, Y, Z)$.

Lemma 14. *Suppose the sequence of laws $\mathcal{L}_n$ and its estimate $\widehat{\mathcal{L}}_n$ satisfy all the assumptions in **Regularity condition**. Then for any given $\alpha \in (0, 1)$, we have for any sequence $M_n \in \mathcal{F}_n$ satisfying $M_n = o_{\mathbb{P}}(1)$,*

$$\mathbb{Q}_{1-\alpha(1+M_n)} \left[n^{1/2} T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z) | X, Y, Z \right] \xrightarrow{\mathbb{P}} z_{1-\alpha}$$

where $T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z)$ is defined as in (59).

To apply Lemma 14 (as well as Lemma 6), we will first verify condition (56)-(58) in **Regularity condition** are satisfied under the assumptions of Theorem 7.

Verification of (56): This is true by assumption (23).

Verification of (57): We verify the condition when $\delta = 2$. When condition (CSE) holds, it suffices to prove

$$\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}_{\widehat{\mathcal{L}}_n} [|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^4 | X, Z] = o_{\mathbb{P}}(1).$$

By Lemma 2, we know

$$\mathbb{E}_{\widehat{\mathcal{L}}_n} [|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^4 | X, Z] = A^{(4)}(\widehat{\theta}_{n,x}(Z_{in})) + 3(A''(\widehat{\theta}_{n,x}(Z_{in})))^2.$$

Since by condition (CSE), $\sup_i |\widehat{\theta}_{n,x}(Z_{in})| = O_{\mathbb{P}}(1)$, we know there exists $\delta > 0$ such that

$$\mathbb{P}[\mathcal{L}] \geq 1 - \delta, \text{ where } \mathcal{L} \equiv \left\{ \sup_i |\widehat{\theta}_{n,x}(Z_{in})| \leq M(\delta) \right\}.$$

Then on the event $\mathcal{L}$, by the smoothness of function A, we have

$$\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(4)}(x)| < \infty,$$
$$\sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x \in [-M(\delta), M(\delta)]} |A^{(2)}(x)| < \infty.$$

Thus we know for any $\delta > 0$,

$$\mathbb{P}\left[\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x\in[-M(\delta),M(\delta)]} |A^{(4)}(x)| < \infty\right] \geq \mathbb{P}[\mathcal{L}] \geq 1-\delta$$
$$\mathbb{P}\left[\sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| \leq \sup_{x\in[-M(\delta),M(\delta)]} |A^{(2)}(x)| < \infty\right] \geq \mathbb{P}[\mathcal{L}] \geq 1-\delta$$

This implies

$$\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| = O_{\mathbb{P}}(1),\ \sup_i |A^{(2)}(\widehat{\theta}_{n,x}(Z_{in}))| = O_{\mathbb{P}}(1).$$

Thus we have

$$\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^4|X,Z] \leq \frac{1}{n}\left(\sup_i |A^{(4)}(\widehat{\theta}_{n,x}(Z_{in}))| + 3\sup_i (A''(\widehat{\theta}_{n,x}(Z_{in})))^2\right)$$
$$= o_{\mathbb{P}}(1).$$

When condition (CCS) holds, it suffices to prove

$$\frac{1}{n^2}\sum_{i=1}^n (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1).$$

This is true by the condition (CCS).

Verification of (58): $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in}|Z_{in}] = A''(\widehat{\theta}(Z_{in})) < \infty$ and $(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 < \infty$ almost surely can be guaranteed respectively by $|\widehat{\theta}(Z_{in})| < \infty$ and $|a_{in}| < \infty$ almost surely in assumption (22). As for $(Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty$, it is true by the integrability of Y_{in}.
Therefore, applying Lemma 14, we have

$$\mathbb{Q}_{1-\alpha(1+M_n)}\left[n^{1/2}T_n^{\mathrm{ndCRT}}(\widetilde{X},X,Y,Z)|X,Y,Z\right] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Moreover, the condition (55) holds for the chosen $T_n(X,Y,Z)$ by condition (64) so that this proves the asymptotic equivalence of $\phi_{n,\alpha}^{\mathrm{aux}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$.

Proof of the equivalence of $\phi_{n,\alpha}^{\mathrm{aux}}, \phi_{n,\alpha}^{\mathrm{dCRT}}$: In order to prove the asymptotic equivalence between $\phi_{n,\alpha}^{\mathrm{dCRT}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$, we apply Lemma 5 with the test statistic $T_n(X,Y,Z)$ to be

$$T_n(X,Y,Z) \equiv \frac{n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)}{\widehat{S}_n^{\mathrm{dCRT}}}$$

and cutoff $C_n(X,Y,Z)$ to be

$$C_n(X,Y,Z) \equiv \mathbb{Q}_{1-\alpha}\left[n^{1/2}T_n^{\mathrm{ndCRT}}(\widetilde{X},X,Y,Z)|X,Y,Z\right].$$

By Lemma 6, we have proved that under the assumptions in Theorem 7,

$$\mathbb{Q}_{1-\alpha}\left[n^{1/2}T_n^{\mathrm{ndCRT}}(\widetilde{X},X,Y,Z)|X,Y,Z\right] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Similarly, the nonaccumulant assumption (55) has been satisfied by (64) so that we have proved the asymptotic equivalence between $\phi_{n,\alpha}^{\mathrm{dCRT}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$. □

E.4 Proof of Corollary 2

Proof of Corollary 2. For any $\varepsilon > 0$,

$$\begin{aligned}
\mathbb{P}_{H_0}[p_{\text{spaCRT}} \le \alpha] &= \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le 0] + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha]] \\
&= \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le 0] + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha], p_{\text{dCRT}}/p_{\text{spaCRT}} \le 1 + \varepsilon] \\
&\quad + \mathbb{P}_{H_0}[p_{\text{spaCRT}} \in (0, \alpha], p_{\text{dCRT}}/p_{\text{spaCRT}} > 1 + \varepsilon] \\
&\le \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le 0] + \mathbb{P}_{H_0}[p_{\text{dCRT}}/p_{\text{spaCRT}} > 1 + \varepsilon] \\
&\quad + \mathbb{P}_{H_0}[p_{\text{dCRT}} \le \alpha(1 + \varepsilon)].
\end{aligned}$$

By the asymptotic validity of dCRT, $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{dCRT}} \le \alpha] \le \alpha$, and conclusion (25) in Theorem 2, we have

$$\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le \alpha] \le \alpha(1 + \varepsilon) + 0 + \lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le 0].$$

By the positivity result $\mathbb{P}_{H_0}[p_{\text{spaCRT}} \le 0] \to 0$ in Theorem 2, we prove

$$\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le \alpha] \le \alpha(1 + \varepsilon).$$

Since $\varepsilon > 0$ is arbitrary, we have $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{spaCRT}} \le \alpha] \le \alpha$. Therefore, spaCRT is asymptotic valid.

□

E.5 Proof of Lemma 14

Proof of Lemma 14. For any given $\varepsilon \in (0, \min\{1/\alpha - 1, 1\}), \eta > 0$, there exists $N(\varepsilon, \eta)$ such that

$$\mathbb{P}[|M_n| > \varepsilon] < \eta, \ \forall n \ge N(\varepsilon, \eta).$$

We will use T_n^{ndCRT} to denote $T_n^{\text{ndCRT}}(\widetilde{X}, X, Y, Z)$. Then consider the $1 - \alpha(1 - \varepsilon)$ and $1 - \alpha(1 + \varepsilon)$ quantiles, we have with probability at least $1 - \eta$, for large enough n, the following is true:

$$\mathbb{Q}_{1-\alpha(1-\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] \ge \mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] \ge \mathbb{Q}_{1-\alpha(1+\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right].$$

Then with probability at least $1 - \eta$, for sufficiently large n, we have

$$|\mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] - z_{1-\alpha}| \le A_n + B_n \tag{70}$$

where

$$A_n \equiv \left|\mathbb{Q}_{1-\alpha(1-\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] - z_{1-\alpha}\right|, \ B_n \equiv \left|\mathbb{Q}_{1-\alpha(1+\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] - z_{1-\alpha}\right|.$$

Applying Lemma 6, we have

$$\mathbb{Q}_{1-\alpha(1-\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] \xrightarrow{\mathbb{P}} z_{1-\alpha(1-\varepsilon)}, \ \mathbb{Q}_{1-\alpha(1+\varepsilon)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] \xrightarrow{\mathbb{P}} z_{1-\alpha(1+\varepsilon)}.$$

Thus for the given ε and sufficiently large n, we have with probability at least $1-\eta$,

$$A_n < \varepsilon + |z_{1-\alpha(1-\varepsilon)} - z_{1-\alpha}|,\ B_n < \varepsilon + |z_{1-\alpha(1+\varepsilon)} - z_{1-\alpha}|.$$

By the continuity of the quantile function of standard normal distribution, we know there exists a universal constant C_α that only depends on α such that

$$|z_{1-\alpha(1+\varepsilon)} - z_{1-\alpha}| < C_\alpha \varepsilon,\ |z_{1-\alpha(1-\varepsilon)} - z_{1-\alpha}| < C_\alpha \varepsilon.$$

Then combining (70), we know with probability at least $1-2\eta$, for sufficiently large n, we have

$$\left|\mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\mathrm{ndCRT}}|X,Y,Z\right] - z_{1-\alpha}\right| \le A_n + B_n < 2C_\alpha\varepsilon + 2\varepsilon.$$

Then since η, ε is arbitrary, we have

$$\mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\mathrm{ndCRT}}(\widetilde{X},X,Y,Z)|X,Y,Z\right] = \mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\mathrm{ndCRT}}|X,Y,Z\right] \xrightarrow{\mathbb{P}} z_{1-\alpha}.$$

Therefore we complete the proof. □

F A special case of binary sampling

We will first present a grand result that states the validity of the spaCRT procedure when $\boldsymbol{X}$ is a binary random variable.

Lemma 15 (Bernoulli sampling)**.** *Suppose $\boldsymbol{X}$ is a binary variable following the natural exponential family model (16) and Assumption 3 holds. Furthermore, suppose*

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1),\ \frac{1}{n}\sum_{i=1}^{n}(\theta(Z_{in}) - \widehat{\theta}_{n,x}(Z_{in}))^2 = o_{\mathbb{P}}(1), \tag{71}$$

and either of the following set of conditions hold:

- ***Fourth moment condition**:*

$$|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \xrightarrow{a.s.} 0,\ |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \xrightarrow{a.s.} 0,\ \forall i \in [n], n \in \mathbb{N}_+; \tag{72}$$

$$\mathbb{P}[|\theta(Z_{in})| < \infty] = 1,\ \sup_n \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}^4] < \infty. \tag{73}$$

- ***Almost sure conditoin:***

$$|\widehat{\theta}_{n,x}(Z_{in})| < \infty, |\widehat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i,n \text{ almost surely}; \tag{74}$$

$$\sup_n \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}^4] < \infty. \tag{75}$$

- ***Bounded condition:***

$$|\widehat{\theta}_{n,x}(Z_{in})| < \infty, |\widehat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i,n \text{ almost surely}; \tag{76}$$

$$\mathbb{P}[\boldsymbol{Y} \in [-S,S]] = 1 \text{ for some } S > 0. \tag{77}$$

Then if condition (24) is true, the conclusion in Theorem 2 holds and

$$(\widehat{S}_n^{\mathrm{dCRT}})^2 = \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2 (X_{in} - \mu_{n,x}(Z_{in}))^2] + o_{\mathbb{P}}(1). \tag{78}$$

Additionally, if the condition (64) is true, the conclusion in Theorem 7 holds.

Lemma 15 will be used to prove Theorems 3-9.

Remark 8. We can see all these assumptions are very mild conditions. In particular, Assumption 3 is a standard non-degeneracy condition. The conditions (71) and (72) only impose the consistency of estimators without the rate condition requirement. This thus gives much flexibility on estimators $\widehat{\mu}_{n,y}(\cdot)$ and $\widehat{\theta}_{n,x}(\cdot)$. Condition (73) is a standard regularity condition. These conditions, together with Assumption 3, are crucial to show the lower bound (23).

We now divide the proof of Lemma 15 into two parts, depending either the **fourth moment condition** or **bounded condition** is used.

F.1 Proof of Lemma 15 with fourth moment condition

Since $\boldsymbol{X}$ is a binary variable and follws model (16), we know $\boldsymbol{X} \mid \boldsymbol{Z} \sim \mathrm{Ber}(\mathrm{expit}(\theta(\boldsymbol{Z})))$. We now verify the conditions in Theorem 2.

Verification of (22): Since $|\theta(Z_{in})| < \infty$ almost surely by condition (73), together with $|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \overset{a.s.}{\to} 0$ in assumption (72), we have

$$|\widehat{\theta}_{n,x}(Z_{in})| \le |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| + |\theta(Z_{in})| < \infty, \text{ almost surely.}$$

We now show $\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2] < \infty$. This implies

$$|Y_{in} - \mu_{n,y}(Z_{in})| < \infty, \forall i \in [n], n \in \mathbb{N}_+ \text{ almost surely.} \tag{79}$$

This is true by using Jensen's inequality (Lemma 10) and Cauchy-Schwarz inequality:

$$\begin{aligned}\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2] &\le 2 \sup_n (\mathbb{E}[Y_{in}^2] + \mathbb{E}[\mu_{n,y}(Z_{in})^2]) \\ &\le 2 \sup_n (\mathbb{E}[Y_{in}^2] + \mathbb{E}[Y_{in}^2]) = 4 \sup_n \mathbb{E}[Y_{in}^2] \le 4 \sqrt{\sup_n \mathbb{E}[Y_{in}^4]} < \infty.\end{aligned}$$

Then by $|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \overset{a.s.}{\to} 0$ in assumption (72), we have

$$|a_{in}| = |Y_{in} - \widehat{\mu}_{n,y}(Z_{in})| \le |Y_{in} - \mu_{n,y}(Z_{in})| + |\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| < \infty$$

almost surely.

Verification of (23): We can write

$$\frac{1}{n}\sum_{i=1} a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in})) = \frac{1}{n}\sum_{i=1}^n a_{in}^2 A''(\theta(Z_{in})) + \frac{1}{n}\sum_{i=1}^n a_{in}(A''(\widehat{\theta}_{n,x}(Z_{in})) - A''(\theta(Z_{in})))$$
$$\equiv T_1 + T_2.$$

We will first prove $T_2 = o_{\mathbb{P}}(1)$. To see this, we notice that $A''(x) = \exp(x)/(1+\exp(x))^2$ and it can be easily checked that $A''(x)$ is a lipschitz function with Lipschitz constant 1. Thus we have

$$|T_2| \le \frac{1}{n}\sum_{i=1}^n a_{in}^2 |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \le \sqrt{\frac{1}{n}\sum_{i=1}^n a_{in}^4}\sqrt{\frac{1}{n}\sum_{i=1}^n |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})|^2}.$$

Thus it suffces to show

$$\frac{1}{n}\sum_{i=1}^n a_{in}^4 = O_{\mathbb{P}}(1),\ \frac{1}{n}\sum_{i=1}^n |\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})|^2 = o_{\mathbb{P}}(1). \tag{80}$$

By assumption (71), it suffcies to show $\frac{1}{n}\sum_{i=1}^n a_{in}^4 = O_{\mathbb{P}}(1)$. In fact, we have

$$\frac{1}{n}\sum_{i=1}^n a_{in}^4 \le \frac{16}{n}\sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^4 + \frac{16}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^4.$$

By the convergence of $\widehat{\mu}_{n,y}(Z_{in})$ in assumption (71), it suffices to show $\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] < \infty$. This is guaranteed by assumption (73) and Jensen's inequality (Lemma 10):

$$\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \le 16 \sup_n (\mathbb{E}[Y_{in}^4] + \mathbb{E}[\mathbb{E}[Y_{in} \mid Z_{in}]^4])$$
$$\le 16 \sup_n (\mathbb{E}[Y_{in}^4] + \mathbb{E}[Y_{in}^4]) = 32 \sup_n \mathbb{E}[Y_{in}^4] < \infty.$$

Therefore, we have proved $T_2 = o_{\mathbb{P}}(1)$ and now we prove $T_1 = \Omega_{\mathbb{P}}(1)$. We first decompose

$$T_1 = \frac{1}{n}\sum_{i=1}^n (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) \equiv \frac{1}{n}\sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) + T_3$$

where

$$T_3 \equiv \frac{1}{n}\sum_{i=1}^n \left\{(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 - (Y_{in} - \mu_{n,y}(Z_{in}))^2\right\} A''(\theta(Z_{in})).$$

Then by the boundedness of A'', we have

$$|T_3| \le \frac{1}{n}\sum_{i=1}^n |\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})||2Y_{in} - \mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in})|$$
$$\le \sqrt{\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in}))^2}\sqrt{\frac{2}{n}\sum_{i=1}^n (Y_{in} - \mu_{n,y}(Z_{in}))^2 + \frac{2}{n}\sum_{i=1}^n a_{in}^2}.$$

We have shown in above that

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in}-\mu_{n,y}(Z_{in}))^2 = O_{\mathbb{P}}(1),\ \frac{1}{n}\sum_{i=1}^{n}a_{in}^2 = O_{\mathbb{P}}(1).$$

Then by assumption (71), we have

$$\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,y}(Z_{in})-\mu_{n,y}(Z_{in}))^2 \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{\mu}_{n,y}(Z_{in})-\mu_{n,y}(Z_{in}))^4} = o_{\mathbb{P}}(1).$$

Thus we have proved $T_3 = o_{\mathbb{P}}(1)$. The final step is to prove

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in}-\mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) = \Omega_{\mathbb{P}}(1).$$

We apply weak law of large numbers to triangular arrays to conclude the proof. In particular, we apply Lemma 7 with $\delta = 1$ so we need to verify

$$\sup_n \mathbb{E}[(Y_{in}-\mu_{n,y}(Z_{in}))^4 (A''(\theta(Z_{in})))^4] < \infty.$$

Since $|A''(x)| \le 1$ for any $x \in \mathbb{R}$, by assumption (73), we have

$$\begin{aligned}\sup_n \mathbb{E}[(Y_{in}-\mu_{n,y}(Z_{in}))^4 (A''(\theta(Z_{in})))^2] &\le \sup_n \mathbb{E}[(Y_{in}-\mu_{n,y}(Z_{in}))^4] \\ &\le 32 \sup_n \mathbb{E}[Y_{in}^4] < \infty.\end{aligned}$$

Therefore, applying Lemma 7 and assumption 3 we obtain

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in}-\mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in})) = o_{\mathbb{P}}(1) + \mathbb{E}[(Y_{in}-\mu_{n,y}(Z_{in}))^2 A''(\theta(Z_{in}))] = \Omega_{\mathbb{P}}(1).$$

This also proves the result (78).

Verification of condition (CCS)**:** Since $\mathbb{P}[\widetilde{X}_{in} \in [-1,1]|\mathcal{F}_n] = 1$ almost surely, it suffices to show

$$\frac{1}{n}\sum_{i=1}^{n}a_{in}^4 = \frac{1}{n}\sum_{i=1}^{n}(Y_{in}-\widehat{\mu}_{n,y}(Z_{in}))^4 = O_{\mathbb{P}}(1).$$

This has been proved in conclusion (80).

F.2 Proof of Lemma 15 with almost sure condition

Checking the proof with **fourth moment condition**, we know the condition (22) is the only part that differs. However, condition (22) has been directly assumed in condition (74). Therefore we complete the proof.

F.3 Proof of Lemma 15 with bounded condition

Verificaitons of conditions (22) and (CCS) are straightforward. As for condition (23), we just note that

$$\sup_n \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \le (2S)^4 = 16S^4 < \infty \quad \text{and} \quad \mathbb{E}[Y_{in}^4] \le S^4 < \infty.$$

Then the proof can go through as that with **fourth moment condition**. Thus we complete the proof.

G Proof of Theorem 3

G.1 Proof of the conclusion in Theorem 2

We need to apply Lemma 15 with **almost sure condition**. In particular, the condition (74) is trivially satisfied by the finiteness of the maximum likelihood estimate $\widehat{\beta}, \widehat{\theta}$. Also condition (75) is true in the low-dimensional setup. Thus it suffices to prove condition (71). We divide the proof into two steps.

Verificaiton of $\frac{1}{n}\sum_{i=1}^n(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1)$ We will prove a stronger result

$$\frac{1}{n}\sum_{i=1}^n(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^4 = O_{\mathbb{P}}(1/n^2). \tag{81}$$

Define the set $C \equiv \{t : |t| \le C_Z(\|\beta\|_1 + \|\theta\|_1) + 1\}$. Then consider the event

$$\mathcal{C} \equiv \{Z_i^\top \widehat{\beta}, Z_i^\top \widehat{\theta} \in C,\ \forall i \in [n]\}.$$

On the event $\mathcal{C}$, we know

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^4 &= \frac{1}{n}\sum_{i=1}^n(A'(Z_i^\top\beta) - A'(Z_i^\top\widehat{\beta}))^4 \\
&\le \frac{1}{n}\sum_{i=1}^n \sup_{t\in C}(A''(t))^4 |Z_i^\top(\beta - \widehat{\beta})|^4 \quad \text{(By mean value theorem)} \\
&\le \sup_{t\in C}(A''(t))^4 \sup_i \|Z_i\|_\infty^4 \|\widehat{\beta} - \beta\|_1^4 \quad \text{(By Hölder's inequality.)} \\
&\le \sup_{t\in C}(A''(t))^4 C_Z^4 \|\widehat{\beta} - \beta\|_1^4 \quad \text{(By Assumption 4)} \\
&= O_{\mathbb{P}}(1/n^2). \quad \text{(By condition (26))}
\end{aligned}$$

Now it suffices to prove $\mathbb{P}[\mathcal{C}] \to 1$. To see this, we compute

$$|Z_i^\top\widehat{\beta}| \le \sup_i \|Z_i\|_\infty \|\widehat{\beta}\|_1 \le C_Z\|\widehat{\beta}\|_1 \le C_Z(\|\beta\|_1 + \|\widehat{\beta} - \beta\|_1)$$

and

$$|Z_i^\top\widehat{\theta}| \le \sup_i \|Z_i\|_\infty \|\widehat{\theta}\|_1 \le C_Z\|\widehat{\theta}\|_1 \le C_Z(\|\theta\|_1 + \|\widehat{\theta} - \theta\|_1).$$

By condition (26), we have $\|\widehat{\beta} - \beta\|_1 = o_{\mathbb{P}}(1)$ and $\|\widehat{\theta} - \theta\|_1 = o_{\mathbb{P}}(1)$. Thus $\mathbb{P}[\mathcal{C}] \to 1$.

Verificaiton of $\frac{1}{n}\sum_{i=1}^n(\theta(Z_{in}) - \widehat{\theta}_{n,x}(Z_{in}))^2 = o_{\mathbb{P}}(1)$ We will also prove a stronger result:

$$\frac{1}{n}\sum_{i=1}^n(\theta(Z_{in}) - \widehat{\theta}_{n,x}(Z_{in}))^2 = \frac{1}{n}\sum_{i=1}^n(Z_i^\top\widehat{\theta} - Z_i^\top\theta)^2 = O_{\mathbb{P}}(1/n).$$

We Hölder's inequality, we have

$$\frac{1}{n}\sum_{i=1}^n(Z_i^\top\widehat{\theta} - Z_i^\top\theta)^2 \le \frac{1}{n}\sum_{i=1}^n\|Z_i\|_\infty^2\|\widehat{\theta} - \theta\|_1^2 \le C_Z^2\|\widehat{\theta} - \theta\|_1^2 = O_{\mathbb{P}}(1/n).$$

G.2 Proof of the asymptotic validity under null

We have verified the conditions for Lemma 15 to hold in section G.1. Thus we know by conclusion (78)

$$(\widehat{S}_n^{\mathrm{dCRT}})^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(Y_i - \mathbb{E}[Y_i|Z_i])^2(X_i - \mathbb{E}[X_i|Z_i])^2] \equiv \sigma_{\mathrm{dCRT}}^2.$$

Thus by Lemma 13, it is sufficient to prove

$$\sqrt{n}T_n^{\mathrm{dCRT}} \xrightarrow{\mathbb{P}} N(0, \sigma_{\mathrm{dCRT}}^2).$$

It is sufficient to prove the following

$$\frac{1}{n}\sum_{i=1}^n(\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 = O_{\mathbb{P}}(1/n) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^n(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 = O_{\mathbb{P}}(1/n).$$

For the first claim, by Cauchy-Schwarz inequality and result (81), we have

$$\frac{1}{n}\sum_{i=1}^n(\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^2 \le \sqrt{\frac{1}{n}\sum_{i=1}^n(\mu_x(Z_i) - \widehat{\mu}_x(Z_i))^4} = O_{\mathbb{P}}(1/n).$$

For the second claim, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 &= \frac{1}{n}\sum_{i=1}^n(\mathrm{expit}(Z_i^\top\theta) - \mathrm{expit}(Z_i^\top\widehat{\theta}))^2 \\
&\le \frac{1}{n}\sum_{i=1}^n(Z_i^\top\theta - Z_i^\top\widehat{\theta})^2 \quad \text{(By Lipschitz continuity of expit)} \\
&\le C_Z\|\widehat{\theta} - \theta\|_1^2 = O_{\mathbb{P}}(1/n). \quad \text{(By Assumption 4 and (26))}
\end{aligned}$$

Therefore we complete the proof.

H Proof of Theorem 4

Let us first explicitly define the estimators $\widehat{\beta}$ and $\widehat{\theta}$.

$$\widehat{\beta}_n = \arg\min_{\beta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{A_y(Z_{in}^\top \beta) - Y_{in} \cdot (Z_{in}^\top \beta)\} + \lambda_n \|\beta\|_1 \right\} \tag{82}$$

and

$$\widehat{\theta}_n = \arg\min_{\theta} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{\log(1 + \exp(Z_{in}^\top \theta)) - X_i \cdot (Z_{in}^\top \theta)\} + \nu_n \|\theta\|_1 \right\}. \tag{83}$$

H.1 Proof of the conclusion in Theorem 2

We first present a lemma which acts as a building block for proving Theorem 4.

Lemma 16. *Suppose all the assumptions in Theorem 4 except for (31) hold. Recall λ_n and ν_n as in models (82) and (83). Then if we choose*

$$\lambda_n = C_\lambda \sqrt{\log(d)/n} \quad \text{and} \quad \nu_n = C_\nu \sqrt{\log(d)/n}$$

for some universal constants C_λ, C_ν, then conclusion in Theorem 2 hold. Furthermore, the variance convergence (78) holds. If additionaly, the condition (64) is true, the conclusion in Theorem 7 holds.

The proof of Lemma 16 will be postponed to section H.4.

H.2 Proof of the asymptotic validity under null

By Lemma 16 and Theorem 7, in order to prove the power and validity of spaCRT, it suffices to show the following conditions are satisfied:

$$\frac{n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z)}{\widehat{S}_n^{\mathrm{dCRT}}} \xrightarrow{d} N(0, 1).$$

The proof is then divided to two parts:

1. We first show the convergence of the test statistic $n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z)$ to a normal distribution;

2. We then show the convergence of the variance $(\widehat{S}_n^{\mathrm{dCRT}})^2$ to the variance of the normal distribution derived.

Then we finish the proof by Slutsky's theorem. In fact, the variance convergence $(\widehat{S}_n^{\mathrm{dCRT}})^2$ has been proved as in Lemma 16. We will dedicate the rest of the proof to show the convergence of $n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z)$. We first decompose $n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z)$ as follows:

$$n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_{in} - \mathbb{E}[X_{in}|Z_{in}])(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}]) + \mathrm{Bias}_1 + \mathrm{Bias}_2 + \mathrm{Bias}_3$$

where

$$\text{Bias}_1 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_{in} - \mathbb{E}[X_{in}|Z_{in}])(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))$$
$$\text{Bias}_2 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbb{E}[X_{in}|Z_{in}] - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])$$
$$\text{Bias}_3 = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\mathbb{E}[X_{in}|Z_{in}] - \widehat{\mu}_{n,x}(Z_{in}))(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})).$$

We will now show that these biases will go to 0 in probability.

Lemma 17 (Bias term convergence)**.** *Suppose all the assumptions in Theorem 4 hold. Then we have*

$$\text{Bias}_1, \text{Bias}_2, \text{Bias}_3 = o_{\mathbb{P}}(1).$$

Now we finish the proof by applying Lemma 12 with $W_{in} = (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))/\sqrt{n}, \mathcal{F}_n = \{\varnothing, \Omega\}$ and $\delta = 2$

$$\frac{\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^4(Y_{in} - \mu_{n,y}(Z_{in}))^4]}{\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2]^2 n} \xrightarrow{\mathbb{P}} 0.$$

converges to 0 in probability. This is true because of the bound

$$\begin{aligned}&\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^4(Y_{in} - \mu_{n,y}(Z_{in}))^4] \le \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \\ &= \mathbb{E}[A_y^{(4)}(Z^\top\theta_n) + 3(A_y''(Z^\top\theta_n))^2] \le \sup_{\|t\|_2 \le C_Z \sup_n \|\theta\|_n} (A_y^{(4)}(t) + 3(A_y''(t))^2)\end{aligned}$$

and $\inf_n \mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2] > 0$ by Assumption 3. Then we have

$$\frac{\sum_{i=1}^n (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in}))}{\sqrt{n\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^2(Y_{in} - \mu_{n,y}(Z_{in}))^2]}} \xrightarrow{d} N(0,1).$$

In other words, by Slutsky's theorem, we have proved

$$\frac{n^{1/2}T_n^{\text{dCRT}}}{\widehat{S}_n^{\text{dCRT}}} \xrightarrow{d} N(0,1).$$

Then we know (55) is satisfied. Thus by Theorem 8, we know

$$\lim_{n\to\infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \lim_{n\to\infty} \mathbb{E}[\phi_{n,\alpha}^{\text{dCRT}}] = \lim_{n\to\infty} \mathbb{P}[\phi_{n,\alpha}^{\text{asy}}] = \alpha.$$

H.3 A strong consistency result for the lasso estimators

Proof of Lemma 16 hinges on a general consistency results proved in (Wainwright, 2019).

Lemma 18 (A modified version of Corollary 9.26 in (Wainwright, 2019)). *Consider the lasso estimators (82) and (83). Suppose assumptions 4-5 and condition (29) hold. If we choose* $\lambda_n = C_\lambda\sqrt{\log(d)/n}$ *and* $\nu_n = C_\nu\sqrt{\log(d)/n}$ *for some universal constants* C_λ, C_ν*, then for any* $\varepsilon > 0$*, there exists* $N(\varepsilon) \in \mathbb{N}$ *such that wheevern* $n \geq N(\varepsilon)$*, we have*

$$\mathbb{P}[\|\widehat{\theta}_n - \theta\|_1 > \varepsilon] \leq 2\exp(-2n^{1-\delta})$$

and

$$\mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > \varepsilon] \leq 2\exp(-2n^{1-\delta}).$$

Consequently, $\|\widehat{\theta}_n - \theta\|_1$ *and* $\|\widehat{\beta}_n - \beta_n\|_1$ *converge to* 0 *almost surely.*

The result is a direct consequence of the concentration inequality for the lasso estimators proved in Corollary 9.26 in (Wainwright, 2019). We now give the proof of the lemma.

Proof of Lemma 18. We need a combination of of Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019) to prove Lemma 18.

Lemma 19 (A combination of Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019)). *Consider the lasso estimators (82) and (83). Suppose assumptions 4-5 holds. If we choose* $\lambda_n = C_\lambda\sqrt{\log(d)/n}$ *and* $\nu_n = C_\nu\sqrt{\log(d)/n}$ *for some universal constants* C_λ, C_ν*, then we have*

$$\mathbb{P}[\|\widehat{\theta}_n - \theta_n\|_1 > C_1 s_{\theta_n}\sqrt{\log(d)/n}] \leq 2\exp(-2\log(d))$$

and

$$\mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > C_2 s_{\beta_n}\sqrt{\log(d)/n}] \leq 2\exp(-2\log(d)).$$

To prove Lemma 18, it is sufficient to show for some $\delta \in (0,1)$,

$$s_{\theta_n}\sqrt{\log(d)/n} = o(1),\ s_{\beta_n}\sqrt{\log(d)/n} = o(1),\ \log(d) \sim n^{1-\delta}.$$

These conditions are satisfied by condition (29). □

H.4 Proof of Lemma 16

We prove the results by applying Lemma 15 combined with the result in Lemma 18. We need to verify conditions (24), (71)-(73).

Verificaiton of (24) Defining $b_{in} \equiv X_{in} - \widehat{\mathbb{E}}[X_{in}|Z_{in}]$, we consider the following decomposition:

$$\begin{aligned} T_n^{\mathrm{dCRT}} &= \frac{1}{n}\sum_{i=1}^n b_{in}(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}]) + \frac{1}{n}\sum_{i=1}^n b_{in}(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})) \\ &\equiv A_n + B_n. \end{aligned}$$

It thus suffices to show $A_n = o_{\mathbb{P}}(1)$ and $B_n = o_{\mathbb{P}}(1)$. We first show $A_n = o_{\mathbb{P}}(1)$ by observing that A_n is just a sample average of conditioally independent random variables on data (X, Z). Since

$$\begin{aligned}
\mathbb{E}[A_n^2|X,Z] &= \frac{1}{n^2}\sum_{i=1}^n b_{in}^2 \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^2|X,Z] \\
&\leq \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^2|X,Z] && (|b_{in}| \leq 1 \text{ almost surely}) \\
&= \frac{1}{n^2}\sum_{i=1}^n A_y''(Z_{in}^\top \beta_n) && (\text{Conditional independence}) \\
&\leq \frac{1}{n}\max_{t\in B} A_y''(t) \quad \text{where} \quad B \equiv \{t \in \mathbb{R} : |t| \leq \sup_n \|\beta_n\|_1 C_Z + 1\}, \\
& && (\text{Compactness of } X, Z \text{ and Hölder's inequality})
\end{aligned}$$

then we know $\mathbb{E}[A_n^2] < \max_{t\in B} A_y''(t)$. By conditional Markov's inequality (Lemma 11): for any $\varepsilon > 0$,

$$\mathbb{P}[|A_n| > \varepsilon|X,Z] \leq \frac{\mathbb{E}[A_n^2|X,Z]}{\varepsilon^2} \leq \frac{\max_{t\in B} A_y''(t)}{n\varepsilon^2}$$

almost surely. Taking expectation on both sides, we have $A_n = o_{\mathbb{P}}(1)$. Now we show $B_n = o_{\mathbb{P}}(1)$. We observe

$$|\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})| = |A_y'(Z_{in}^\top \beta_n) - A_y'(Z_{in}^\top \widehat{\beta}_n)|. \tag{84}$$

On the event

$$E_n \equiv \left\{ Z_{in}^\top \beta_n, Z_{in}^\top \widehat{\beta}_n \in B \text{ for any } i \in [n] \right\}, \tag{85}$$

we know

$$\begin{aligned}
|B_n| &\leq \frac{1}{n}\sum_{i=1}^n |b_{in}(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && (\text{Traingle inequality}) \\
&\leq \frac{1}{n}\sum_{i=1}^n |(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && (|b_{in}| \leq 1 \text{ almost surely}) \\
&= \frac{1}{n}\sum_{i=1}^n |A_y'(Z_{in}^\top \beta_n) - A_y'(Z_{in}^\top \widehat{\beta}_n)| && (\text{By result (84)}) \\
&\leq \frac{1}{n}\sum_{i=1}^n \max_{t\in B} A_y''(t)|Z_{in}^\top(\widehat{\beta}_n - \beta_n)| && (\text{Mean value theorem}) \\
&\leq \max_{t\in B} A_y''(t) C_Z \|\widehat{\beta}_n - \beta_n\|_1. && (\text{Hölder's inequality})
\end{aligned}$$

By Lemma 18, we know $\|\widehat{\beta}_n - \beta_n\|_1$ coverge to 0 almost surely. Thus it suffices to show $\mathbb{P}[E_n] \to 0$. In fact, we can prove a stronger result.

Lemma 20. *Suppose the assumptions in Theorem 16 hold. Then*

$$\mathbb{P}[E_n^c \text{ happens infinitely often}] = 0.$$

Proof of Lemma 20 can be found in section H.6. With Lemma 20, we conclude the verificaiton for (24).

Verificaiton of (71) We first show $(1/n)\sum_{i=1}^n(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 = o_{\mathbb{P}}(1)$. We know one the event E_n (defined in (85)),

$$|A_y'(Z_{in}^\top \widehat{\beta}_n)| \le \max_{t\in B} A_y''(t)(C_Z\|\widehat{\beta}_n - \beta_n\|_1).$$

Thus we know on the event E_n,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^4 &= \frac{1}{n}\sum_{i=1}^n(A'(Z_{in}^\top\beta_n) - A'(Z_{in}^\top\widehat{\beta}_n))^4 \\
&\le \sup_{t\in B}(A''(t))^4\frac{1}{n}\sum_{i=1}^n|Z_{in}^\top(\widehat{\beta}_n - \beta_n)|^4 \\
&\qquad\qquad \text{(By mean value theorem)} \\
&\le \sup_{t\in B}(A''(t))^4 C_Z^4\|\widehat{\beta}_n - \beta_n\|_1^4. \quad \text{(By Hölder's inequality)}
\end{aligned}$$

Thus by Lemma 18, we know $\|\widehat{\beta}_n - \beta_n\|_1 \xrightarrow{\mathbb{P}} 0$ under the given assumption.

Verification of (72) We first show $|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| \to 0$ almost surely. It suffices to prove these claims on the event E_n by Lemma 20. In fact, on E_n, using again the mean value theorem, we can prove

$$\begin{aligned}
|\widehat{\mu}_{n,y}(Z_{in}) - \mu_{n,y}(Z_{in})| &= |A'(Z_{in}^\top\widehat{\beta}_n) - A'(Z_{in}^\top\beta_n)| \\
&\le \sup_{t\in B}|A''(t)||Z_{in}^\top(\widehat{\beta}_n - \beta_n)| \\
&\le \sup_{t\in B}|A''(t)|C_Z\|\widehat{\beta}_n - \beta_n\|_1 \to 0
\end{aligned}$$

The last convergence holds almost surely by Lemma 18. Now we prove the claim that $|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \to 0$ almost surely. We can bound, using Hölder's inequality,

$$|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})| \le |Z_{in}^\top\widehat{\theta}_n - Z_{in}^\top\theta_n| \le C_Z\|\widehat{\theta}_n - \theta_n\|_1 \to 0$$

almost surley by Lemma 18. This concludes the verification of (72).

Verification of (73) We first show $|\theta(Z_{in})| < \infty$ almost surely. This is because of the following bound:

$$|\theta(Z_{in})| = |Z_{in}^\top\theta_n| \le \|Z_{in}\|_\infty\|\theta_n\|_1 \le C_Z\sup_n\|\theta_n\|_1 < \infty.$$

Noe we prove the claim $\sup_n \mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Y}^4] < \infty$. By the representation $\mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}])^4] = \mathbb{E}[A_y^{(4)}(\boldsymbol{Z}^\top\theta_n) + 3(A_y''(\boldsymbol{Z}^\top\theta_n))^2]$, we can bound

$$\begin{aligned}
\mathbb{E}[\boldsymbol{Y}^4] &\le 16\mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}])^4] + 16\mathbb{E}[(\mathbb{E}[\boldsymbol{Y}|\boldsymbol{Z}])^4] \\
&\le 16\mathbb{E}[A_y^{(4)}(Z^\top\theta_n)] + 48\mathbb{E}[(A_y''(Z^\top\theta_n))^2] + 16\mathbb{E}[(A_y'(Z^\top\theta_n))^4].
\end{aligned}$$

It suffices to prove there exits a universal constant C_M such that the following three statements:

$$\max\{\sup_n A_y^{(4)}(Z^\top\theta_n), \sup_n A_y''(Z^\top\theta_n), \sup_n A_y'(Z^\top\theta_n)\} \le C_M < \infty.$$

This can be shown by noticing

$$\sup_n A_y^{(4)}(Z^\top\theta_n) \le \sup_{t\in B}|A_y^{(4)}(t)|,\ \sup_n A_y''(Z^\top\theta_n) \le \sup_{t\in B}|A_y''(t)|$$

and

$$\sup_n A_y'(Z^\top\theta_n) \le \sup_{t\in B}|A_y'(t)|.$$

Thus C_M can be chosen to be $\max\{\sup_{t\in B}|A_y^{(4)}(t)|, \sup_{t\in B}|A_y''(t)|, \sup_{t\in B}|A_y'(t)|\}$. This concludes the verification of (73).

H.5 Proof of Lemma 17

For Bias_1, by conditional Markov's inequality (Lemma 11) we have

$$\mathbb{P}[\text{Bias}_1 \ge \varepsilon|Y,Z] \le \frac{\mathbb{E}[\text{Bias}_1^2|Y,Z]}{\varepsilon^2}.$$

Then we have

$$\begin{aligned}\mathbb{E}[\text{Bias}_1^2|Y,Z] &= \frac{1}{n}\sum_{i=1}^n (\text{expit})'(Z_{in}^\top\theta_n)(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^2 \\ &\le \frac{1}{n}\sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^2. \quad \text{(Derivative of expit}(\cdot)\text{ is bounded.)}\end{aligned}$$

Similarly, for Bias_2, we have

$$\begin{aligned}\mathbb{P}[\text{Bias}_2 \ge \varepsilon|X,Z] &\le \frac{\mathbb{E}[\text{Bias}_2^2|X,Z]}{\varepsilon^2} \\ &\le \frac{1}{\varepsilon^2 n}\sum_{i=1}^n \mathbb{E}[(Y_{in} - \mathbb{E}[Y_{in}|Z_{in}])^2|Z_{in}](\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2 \\ &= \frac{1}{\varepsilon^2 n}\sum_{i=1}^n A_y''(Z_{in}^\top\theta_n)(\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2 \\ &\le \frac{\sup_{\|t\|_2\le C_Z\|\theta_n\|_1} A_y''(t)}{\varepsilon^2 n}\sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2. \\ &\qquad\qquad \text{(By Hölder's inequality)}\end{aligned}$$

For Bias_3, we use Cauchy-Schwarz inequality to bound

$$|\text{Bias}_3| \le \sqrt{n}\sqrt{\frac{1}{n}\sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2}\sqrt{\frac{1}{n}\sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^2}.$$

Thus combining above observations, it suffice to prove

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,x}(Z_{in})-\widehat{\mu}_{n,x}(Z_{in}))^2=o_p(1),\quad \frac{1}{n}\sum_{i=1}^{n}(\mu_{n,y}(Z_{in})-\widehat{\mu}_{n,y}(Z_{in}))^2=o_p(1) \tag{86}$$

and

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,x}(Z_{in})-\widehat{\mu}_{n,x}(Z_{in}))^2\right)\cdot\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,y}(Z_{in})-\widehat{\mu}_{n,y}(Z_{in}))^2\right)=o_p(1/n). \tag{87}$$

We compute

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,x}(Z_{in})-\widehat{\mu}_{n,x}(Z_{in}))^2 &\le \frac{1}{n}\sum_{i=1}^{n}(Z_{in}^\top\hat{\theta}_n-Z_{in}^\top\theta_n)^2 && \text{(Lipschitz property)}\\ &\le C_Z\|\hat{\theta}_n-\theta_n\|_1^2 && \text{(Hölder's inequality)}\end{aligned}$$

Then by Lemma 18, we know $\|\hat{\theta}_n-\theta_n\|_1^2=O_p(s_{\theta_n}^2\log(d)/n)$. Thus we have

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,x}(Z_{in})-\widehat{\mu}_{n,x}(Z_{in}))^2=O_p(s_{\theta_n}^2\log(d)/n).$$

We can compute

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,y}(Z_{in})-\widehat{\mu}_{n,y}(Z_{in}))^2\\ &=\frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}[Y_{in}|Z_{in}]-A_y'(Z_{in}^\top\widehat{\beta}_n)\text{expit}(Z_{in}^\top\widehat{\theta}_n)-A_y'(Z_{in}^\top\widehat{\beta}_n)(1-\text{expit}(Z_{in}^\top\widehat{\theta}_n)))^2\\ &\le\frac{2}{n}\sum_{i=1}^{n}(A_y'(Z_{in}^\top\beta_n)-A_y'(Z_{in}^\top\widehat{\beta}_n))^2+\frac{2}{n}\sum_{i=1}^{n}(A_y'(Z_{in}^\top\beta_n)-A_y'(Z_{in}^\top\widehat{\beta}_n))^2\\ &\le\frac{2}{n}\sum_{i=1}^{n}\sup_{t\in B}(A_y''(t))^2(Z_{in}^\top\widehat{\beta}_n-Z_{in}^\top\beta_n)^2+\frac{2}{n}\sum_{i=1}^{n}\sup_{t\in B}(A_y''(t))^2(Z_{in}^\top\widehat{\beta}_n-Z_{in}^\top\beta_n)^2\\ &=O(\|\widehat{\beta}_n-\beta_n\|_1^2)\\ &=O_p(s_{\beta_n}^2\log(d)/n).\end{aligned}$$

where the last inequality is due to mean value theorem and B is defined as in the event (85). Therefore combining the above results and conditions (29) and (31), we know claims (86) and (87) hold so that we finish the proof.

H.6 Proof of Lemma 20

Proof of Lemma 20. First consider $Z_{in}^\top\beta_n$. By Hölder's iequality, we have

$$|Z_{in}^\top\beta_n|\le C_Z\sup_n\|\beta_n\|_1\le C_Z\sup_n\|\beta_n\|_n+1.$$

Now we consider $Z_{in}^\top \widehat{\beta}_n$. Similarly, we have

$$|Z_{in}^\top \widehat{\beta}_n| \le C_Z \|\widehat{\beta}_n - \beta_n\|_1 + C_Z \|\beta_n\|_1.$$

Since $\|\widehat{\beta}_n - \beta_n\|_1 \to 0$ almost surely by Lemma 18, we know

$$\begin{aligned}\mathbb{P}[Z_{in}^\top \widehat{\beta}_n \in B, \forall i \in [n]] &\ge \mathbb{P}[C_Z \|\widehat{\beta}_n - \beta_n\|_1 + C_Z \|\beta_n\|_1 \le C_Z \sup_n \|\beta_n\|_1 + 1] \\ &\ge \mathbb{P}[C_Z \|\widehat{\beta}_n - \beta_n\|_1 \le |1] \ge \mathbb{P}[C_Z \|\widehat{\beta}_n - \beta_n\|_1 \le 1].\end{aligned}$$

Thus we know

$$\mathbb{P}[E_n^c] \le 2\mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > 1/C_Z].$$

By Lemma 18, we know there exists $N(1/C_Z)$ such that for any $n \ge N(1/C_Z)$,

$$\mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > 1/C_Z] \le \exp(-2n^{1-\delta}).$$

Then we have

$$\sum_{n=1}^{\infty} \mathbb{P}[E_n^c] = \sum_{n=1}^{\infty} \mathbb{P}[\|\widehat{\beta}_n - \beta_n\|_1 > 1/C_Z] \le N(1/C_Z) + 2\sum_{n=1}^{\infty} \exp(-2n^{1-\delta}) < \infty.$$

This concludes the proof. □

I Proof of Theorem 9

We divide the proof to two parts: proof of the approximation accuracy conclusion in Theorem 2 and proof of the asymptotic validity of spaCRT under the null hypothesis. They will be presented in section I.2 and section I.3 respectively. We first present a general consistency result on the KRR estimator which will be used in both proofs.

I.1 A general upper bound on the KRR estimator

Before proceeding to the proof, we present a key lemma which states the finite-sample bound on the in-sample mean square error of the KRR estimator. The statement is a combination of Lemma 30 and 31 in Shah and Peters (2020).

Lemma 21 (KRR finite-sample bound)**.** *Let $Z_1, \dots, Z_n \in \mathbb{R}^d$ be independently and identically distributed observations and suppose for any $j \ne i$,*

$$X_i = f(Z_i) + \varepsilon_i \quad \text{where} \quad \mathrm{Var}[\varepsilon_i] < \sigma^2 \text{ and } \mathrm{Cov}[\varepsilon_i, \varepsilon_i] = 0$$

for $f \in \mathcal{H}$ for some RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ with reproducing kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Consider the kernel matrix $K \in \mathbb{R}^{n\times n}$ have ijth $K_{ij} = k(Z_i, Z_j)/n$ and denote the eigvenvalue in the eigen-expansion (92) of kernel k as $\kappa_1 \ge \kappa_2 \ge \cdots \ge \kappa_n \ge 0$, then, for any $\lambda > 0$, the regularization estimator takes the form

$$\widehat{f}_\lambda \equiv K(K + \lambda I)^{-1} X \quad \text{where} \quad X = (X_1, \dots, X_n)^\top \in \mathbb{R}^n,$$

and satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[(f(Z_i) - \widehat{f}_\lambda(Z_i))^2 | Z_1, \dots, Z_n] \le \frac{\sigma^2}{\lambda}\frac{1}{n}\sum_{i=1}^{n} \min\{\kappa_i/4, \lambda\} + \|f\|_{\mathcal{H}}^2 \frac{\lambda}{4}.$$

A direct corollary for Lemma 21 is the convergence of mean square error as a consequence of conditional Markov's inequality (Lemma 11).

Corollary 3 (Convergence of KRR estimator)**.** *Suppose the conditions in Lemma 21 hold. Then if condition (98) holds, we have*

$$\frac{1}{n}\sum_{i=1}^{n}(f(Z_i)-\widehat{f}_\lambda(Z_i))^2 = o_{\mathbb{P}}(1).$$

I.2 Proof of the conclusion in Theorem 2

We will verify the conditions in the version of Lemma 15 with **bounded condition**. In fact, condition (77) is clearly satisfied by the boundedness of $\boldsymbol{Y}$ (condition (95)) and we now prove the other conditions in the lemma.

Verification of condition (71) We first show, by Corollary 3 and condition (97),

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^4 \le \frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^2(S+\sup_{z\in\mathbb{R}^d}|\widehat{\mu}_y(z)|)^2 = o_{\mathbb{P}}(1).$$

Then we show

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\theta(Z_{in})-\widehat{\theta}_{n,x}(Z_{in}))^2 &= \frac{1}{n}\sum_{i=1}^{n}(Z_i^\top\widehat{\theta}-Z_i^\top\theta)^2 \\
&\le \sup_i\|Z_i\|_\infty\|\widehat{\theta}-\theta\|_1 && \text{(By Hölder's inequality)} \\
&\le C_Z\|\widehat{\theta}-\theta\|_1 && \text{(By Assumption 4)} \\
&= o_{\mathbb{P}}(1). && \text{(By condition (96))}
\end{aligned}$$

This completes the verification of condition (71).

Verificaiton of condition (76) The $\widehat{\theta}_{n,x}(Z_{in}) = Z_i^\top\widehat{\theta}$ is finite almost surely by the definition of maximum likelihood estimator. We will show that $|\widehat{\mu}_y(Z_i)|<\infty$ almost surely for any i. This is obvious by noticing the KRR estimators:

$$\widehat{\mu}_y(Z_i) = K_{Z_i}(K+\lambda_n I)^{-1}Y, \quad \text{where} \quad K_{Z_i} = (K_{1i},\ldots,K_{ni}).$$

Then by the finiteness and positivity of λ_n and the positive semidefiniteness of the kernel matrix K, we know the claim is true.

Verificaiton of condition (24) We prove a stronger result:

$$\sqrt{n}T_n^{\mathrm{dCRT}} \xrightarrow{d} N(0,\sigma_{\mathrm{dCRT}}^2) \quad \text{where} \quad \sigma_{\mathrm{dCRT}}^2 = \mathbb{E}[(X_i-\mu_x(Z_i))^2(Y_i-\mu_y(Z_i))^2]. \tag{88}$$

By Lemma 13, it suffices to show the following

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2 = o_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^2 = o_{\mathbb{P}}(1) \\
&\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i)-\widehat{\mu}_y(Z_i))^2\right) = o_{\mathbb{P}}(1/n),
\end{aligned}$$

and

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu_x(Z_i))(Y_i-\mu_y(Z_i)) \xrightarrow{d} N(0,\sigma^2_{\text{dCRT}}).$$

The last claim is justfied by a application of central limit Theorem under the existence of the second moment. We now focus on proving the other claims. For the first claim, we have

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2 \le \sup_i \|Z_i\|_\infty \|\widehat{\theta}-\theta\|_1 \le C_Z\|\widehat{\theta}-\theta\|_1 = o_{\mathbb{P}}(1).$$

For the second claim, we know it is true by Corollary 3. Next, we prove the third claim. This is obvious by the above arugments and condition (96).

I.3 Proof of the asymptotic validity under null

By Theorem 8, we just need to show that

$$\frac{\sqrt{n}T_n^{\text{dCRT}}}{\widehat{S}_n^{\text{dCRT}}} \xrightarrow{d} N(0,1)$$

Note that this statement implies condition (55) holds. We have proved the conditions required in Lemma 15 in section I.2 so we know by the conclusion (78) that

$$(\widehat{S}_n^{\text{dCRT}})^2 \xrightarrow{\mathbb{P}} \sigma^2_{\text{dCRT}} \equiv \mathbb{E}[(X_i-\mu_x(Z_i))^2(Y_i-\mu_y(Z_i))^2].$$

We have proved the conclusiogn (88) so by Slutsky's theorem, the desired claim is true.

J A preliminary introduction to HMM

J.1 Generating genetic variable from multinomial HMM

Consider matrix $X \in \mathbb{R}^{n\times p}$ with i.i.d. rows following the law of $\boldsymbol{X} \in \{0,1\}^d$. Each variable of $\boldsymbol{X}$, $\boldsymbol{X}_j$, can be thought as a copy inherited from either paternal side or maternal side. Multinomial HMM (mHMM) is a mathematical model to model $\boldsymbol{X}_j$ because of its probabilistic structure mimicking the hereditary nature. Before introducing the mHHM, let us first define a multinomial Markov chain (mMC) as follows:

Definition 1 (multinomial Makrov chain)**.** *We say a random variable $\boldsymbol{U} \in \{0,1,\dots,K\}^d$ follows a multinomial Marko Chain distribution, $\boldsymbol{U} \sim mMC(q,Q)$, if*

$$\boldsymbol{U}_1 \sim q \quad \textit{and} \quad \boldsymbol{U}_j|\boldsymbol{U}_{j-1} \sim Q(\cdot|\boldsymbol{U}_{j-1}),$$

where q is some multinomial distribution supported on $\{0,1,\dots,K\}$ and $Q(\cdot|\cdot) \in \mathbb{R}^{K\times K}$ is a transition matrix so that $Q(\cdot|\boldsymbol{U}_{j-1}=u)$ is a transition distribution given the observation $\boldsymbol{U}_{j-1}$ is u.

Then we consider the following definition for mHMM:

Definition 2 (mHMM). *Consider a p-dimensional random variable* $\boldsymbol{X} \in \{0,1\}^d$. *We say* $\boldsymbol{X} \sim mHMM(\boldsymbol{U}(q,Q),e)$ *is an observation from a multinomial hidden Markov model with a latent multinomial Markov chain* $\boldsymbol{U} \equiv (\boldsymbol{U}_1,\ldots,\boldsymbol{U}_p)^\top \sim mMC(q,Q)$ *and a multinomial emission distribution* $\boldsymbol{X}_j|\boldsymbol{U}_j \sim e$. *In particular, we denote* $\mathbb{P}[\boldsymbol{X}_j = x_j|\boldsymbol{U}_j = u_j] \equiv e(x_j|u_j)$.

In particular, we assume $\boldsymbol{X} \equiv (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p)^\top$ is generated from a mHMM:

$$\boldsymbol{X} \sim \text{mHMM}(q,\boldsymbol{U}(Q,e)).$$

The graphical illustration of the latent DGP for observed genetic variable $\boldsymbol{X} = (\boldsymbol{X}_1,\ldots,\boldsymbol{X}_p)^\top$ is shown in Figure 5.

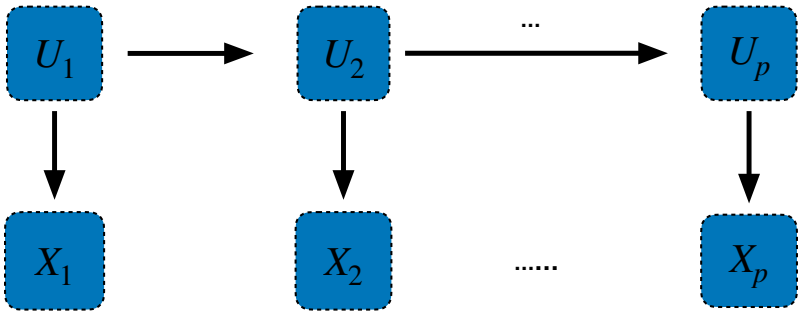

Figure 5: Graphical illustration of a multinomial hidden Markov model. Dashed rounded rectangles: unobserved variables.

The following identities can be easily obtained:

$$\mathbb{P}[\boldsymbol{X}_j,\boldsymbol{U}_{j+1}|\boldsymbol{U}_j] = \mathbb{P}[\boldsymbol{X}_j|\boldsymbol{U}_j]\mathbb{P}[\boldsymbol{U}_{j+1}|\boldsymbol{U}_j]; \tag{CI}$$

$$\mathbb{P}[\boldsymbol{X}_{j-1},\boldsymbol{U}_j|\boldsymbol{U}_{1:(j-1)},\boldsymbol{X}_{1:(j-2)}] = \mathbb{P}[\boldsymbol{X}_{j-1},\boldsymbol{U}_j|\boldsymbol{U}_{j-1}], \tag{Markov}$$

where we use $\boldsymbol{U}_{1:j}$ to denote $(\boldsymbol{U}_1,\ldots,\boldsymbol{U}_j)^\top \in \mathbb{R}^j$ and same notation is applied to $\boldsymbol{X}_{1:j}$.

J.2 Computing conditional distribution-related quantities

Define $x_{-j} \equiv (x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)^\top \in \mathbb{R}^{p-1}$. After the joint distribution $\boldsymbol{X}$ has been estimated from fitting the mHMMs, we can compute the conditional distribution $\boldsymbol{X}_j|\boldsymbol{X}_{-j}$ for any $j \in [p]$ with the estimated parameters. We will dedicate this section to discussing how the conditional expectation, conditional cumulant generating function (CCGF) and its derivatives, required by the spaCRT method, can be easily computed. We first boil these quantities down to the conditional probability $\mathbb{P}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = x_{-j}]$. For the ease of notation, we now define

$$p(x_j,x_{-j}) \equiv \mathbb{P}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = x_{-j}].$$

1. **Conditional expectation:** For any given $x_{-j} \in \mathbb{R}^{p-1}$, we can compute

$$\mathbb{E}[\boldsymbol{X}_j | \boldsymbol{X}_{-j} = x_{-j}] = p(1, x_{-j}) + 2 \cdot p(2, x_{-j}).$$

2. **Function value and derivatives of CCGF:** spaCRT method requires the knowledge of the CCGF function and its derivatives up to second order. Defining $D(t, x_{-j}) \equiv \mathbb{E}[\exp(t\boldsymbol{X}_j)|\boldsymbol{X}_{-j} = x_{-j}]$, we can compute

$$D(t, x_{-j}) = p(0, x_{-j}) + \exp(t)p(1, x_{-j}) + \exp(2t)p(2, x_{-j}).$$

Thus we can compute the CCGF value:

$$K(t, x_{-j}) \equiv \log \mathbb{E}[\exp(t\boldsymbol{X}_j)|\boldsymbol{X}_{-j} = x_{-j}] = \log(D(t, x_{-j})) \quad \forall t \in \mathbb{R}.$$

Then we compute the first derivative of CCGF:

$$\nabla_t K(t, x_{-j}) = \frac{\exp(t)p(1, x_{-j}) + 2\exp(2t)p(2, x_{-j})}{D(t, x_{-j})}$$

and the second derivative of CCGF:

$$\nabla_t^2 K(t, x_{-j}) = \frac{\exp(t)p(1, x_{-j}) + 4\exp(2t)p(2, x_{-j})}{D(t, x_{-j})} - \big[\nabla_t K(t, x_{-j})\big]^2.$$

Thus from the above computation, we know it is sufficient to compute the conditional probability $p(x_j, x_{-j})$. In fact, the following Proposition shows that it can be computed iteratively. To state the Proposition, we need to introduce necessary notation. For $u, \bar{u} \in \{0, 1, \ldots, K\}$ and $x \in \{0, 1\}^d$, we define

$$A_j(u, x) \equiv \mathbb{P}[\boldsymbol{X}_{1:(j-1)} = x_{1:(j-1)}, \boldsymbol{U}_j = u], \ A_1(u, x) \equiv \mathbb{P}[\boldsymbol{U}_1 = u]; \tag{89}$$

$$B_j(u, x) \equiv \mathbb{P}[\boldsymbol{X}_{(j+1):p} = x_{(j+1):p} | \boldsymbol{U}_j = u], \ B_p(u, x) \equiv 1. \tag{90}$$

Proposition 2 (Iterative computation of $p(x_j, x_{-j})$)**.** *Suppose $\boldsymbol{X} \sim mHMM(\boldsymbol{U}(q, Q), e)$. Then we have*

$$p(x_j, x_{-j}) = \frac{\sum_{\bar{u} \in \{0,1,\ldots,K\}} e(x_j|\bar{u}) \cdot A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}{\sum_{\bar{u} \in \{0,1,\ldots,K\}} A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}. \tag{91}$$

In particular, A_j and B_j can be computed in the following recursive manner: for any $u \in \{0, 1, \ldots, K\}$ and $x \in \{0, 1\}^d$,

$$A_j(z, x) = \sum_{\bar{u} \in \{0,1,\ldots,K\}} A_{j-1}(\bar{u}, x) \cdot r(x_{j-1}|\bar{u}) \cdot Q(u|\bar{u});$$

$$B_j(u, x) = \sum_{\bar{u} \in \{0,1,\ldots,K\}} B_{j+1}(\bar{u}, x) \cdot r(x_{j+1}|\bar{u}) \cdot Q(\bar{u}|u).$$

The proof can be found in Appendix J.3.

J.3 Proof of Proposition 2

Proof of conclusion (91) First, let us consider the following marginalization:

$$\begin{aligned}
&\mathbb{P}[\boldsymbol{X}_j = x_j | \boldsymbol{X}_{-j} = x_{-j}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_j = x_j | \boldsymbol{X}_{-j} = x_{-j}, \boldsymbol{U}_j = \bar{u}] \mathbb{P}[\boldsymbol{U}_j = \bar{u} | \boldsymbol{X}_{-j} = x_{-j}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_j = x_j | \boldsymbol{U}_j = \bar{u}] \mathbb{P}[\boldsymbol{U}_j = \bar{u} | \boldsymbol{X}_{-j} = x_{-j}] \qquad \text{(Markov property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} e(x_j|\bar{u}) \mathbb{P}[\boldsymbol{U}_j = \bar{u} | \boldsymbol{X}_{-j} = x_{-j}].
\end{aligned}$$

Now we compute the conditioanl probability $\mathbb{P}[\boldsymbol{U}_j = u | \boldsymbol{X}_{-j} = x_{-j}]$. In particular, we consider the following decomposition

$$\mathbb{P}[\boldsymbol{U}_j = u | \boldsymbol{X}_{-j} = x_{-j}] = \frac{\mathbb{P}[\boldsymbol{U}_j = u, \boldsymbol{X}_{-j} = x_{-j}]}{\mathbb{P}[\boldsymbol{X}_{-j} = x_{-j}]} = \frac{\mathbb{P}[\boldsymbol{U}_j = u, \boldsymbol{X}_{-j} = x_{-j}]}{\sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_j = \bar{u}, \boldsymbol{X}_{-j} = x_{-j}]}.$$

Thus we have

$$\mathbb{P}[\boldsymbol{X}_j = x_j | \boldsymbol{X}_{-j} = x_{-j}] = \frac{\sum_{\bar{u}\in\{0,1,\ldots,K\}} e(x_j|\bar{u}) \mathbb{P}[\boldsymbol{U}_j = \bar{u}, \boldsymbol{X}_{-j} = x_{-j}]}{\sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_j = \bar{u}, \boldsymbol{X}_{-j} = x_{-j}]}$$

so that we only need to compute the probability $\mathbb{P}[\boldsymbol{U}_j = u, \boldsymbol{X}_{-j} = x_{-j}]$. Further, we can do the following calculation

$$\begin{aligned}
\mathbb{P}[\boldsymbol{U}_j = u, \boldsymbol{X}_{-j} = x_{-j}] &= \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u] \\
&= \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_j = u] \quad \text{(Markov property)} \\
&= A_j(u, x) \cdot B_j(u, x).
\end{aligned}$$

The detailed derivation of $A_j(u, x)$ and $B_j(u, x)$ will be present in the next two sections.

Computing $A_j(u, x)$ using forward algorithm We use prove-by-induction to derive $A_j(u, x)$ and the induction is on index j. Since $A_1(u, x) = \mathbb{P}[\boldsymbol{U}_1 = u]$ and we can compute:

$$\begin{aligned}
A_2(u, x) &= \mathbb{P}[\boldsymbol{X}_1 = x_1, \boldsymbol{U}_2 = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_1 = \bar{u}] \mathbb{P}[\boldsymbol{X}_1 = x_1, \boldsymbol{Z}_2 = z | \boldsymbol{U}_1 = \bar{u}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_1 = \bar{u}] \mathbb{P}[\boldsymbol{X}_1 = x_1 | \boldsymbol{U}_1 = \bar{u}] \mathbb{P}[\boldsymbol{U}_2 = u | \boldsymbol{U}_1 = \bar{u}] \quad \text{(CI property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} A_1(\bar{u}, x) \cdot e(x_1|\bar{u}) \cdot Q(u|\bar{u}). \qquad \text{(Definition (89))}
\end{aligned}$$

Then we can easily show that

$$\begin{aligned}
A_j(u,x) &= \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{Z}_j = z | \boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u | \boldsymbol{U}_{j-1} = \bar{u}] \qquad \text{(Markov property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} A_{j-1}(\bar{u}, x) \cdot e(x_{j-1}|\bar{u}) \cdot Q(u|\bar{u}). \qquad \text{(Definition (89))}
\end{aligned}$$

With iterative computation, one can obtain the probability $A_j(u,x)$ for any $j \in [d]$ and $u \in \{0,1,\ldots,K\}$ and $x \in \{0,1\}^d$.

Computing $B_j(u)$ using backward algorithm Using $B_p(u,x) = 1$ for any u, we can compute

$$\begin{aligned}
B_{p-1}(u,x) &= \mathbb{P}[\boldsymbol{X}_p = x_p | \boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p, \boldsymbol{U}_p = \bar{u} | \boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p | \boldsymbol{U}_p = \bar{u}, \boldsymbol{U}_{p-1} = u] \mathbb{P}[\boldsymbol{U}_p = \bar{u} | \boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p | \boldsymbol{U}_p = \bar{u}] \mathbb{P}[\boldsymbol{U}_p = \bar{u} | \boldsymbol{U}_{p-1} = u] \quad \text{(Markov property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} B_p(\bar{u}) \cdot r(x_p|\bar{u}) \cdot s(\bar{u}|u). \qquad \text{(Definition (90))}
\end{aligned}$$

By induction, we can compute

$$\begin{aligned}
B_j(u,x) &= \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p, \boldsymbol{U}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+2} = x_{j+2}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_{j+1} = \bar{u}, \boldsymbol{U}_j = u, \boldsymbol{X}_{j+1} = x_{j+1}] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \boldsymbol{Z}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+2} = x_{j+2}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_{j+1} = \bar{u}] \qquad \text{(Markov property)} \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1} | \boldsymbol{U}_{j+1} = \bar{u}] \cdot \mathbb{P}[\boldsymbol{U}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \qquad \text{(CI property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} B_{j+1}(\bar{u}, x) \cdot e(x_{j+1}|\bar{u}) \cdot Q(\bar{u}|u). \qquad \text{(Definition (90))}
\end{aligned}$$

With iterative computation, one can obtain the probability $B_j(u,x)$ for any $j \in [d]$ and $u \in \{0,1,\ldots,K\}$ and $x \in \{0,1\}^d$.

Final form: combining $A_j(u,x)$ and $B_j(u,x)$ Now we compute the probability $\mathbb{P}[\boldsymbol{U}_j = u_j, \boldsymbol{X}_{-j} = x_{-j}]$ using the derivation in the above sections. In partiuclar, fixing a set of value x_{-j}, we compute

$$\mathbb{P}[\boldsymbol{Z}_j = z, \boldsymbol{X}_{-j} = x_{-j}] = \frac{A_j(z,x) \cdot B_j(z,x)}{\sum_{\bar{z} \in \{0,1,\ldots,K\}} A_j(\bar{u},x) \cdot B_j(\bar{u},x)} \quad \forall j \in [d],$$

where $A_j(u,x), B_j(u,x)$ can be computed iteratively as before.

K Additional simulation details in Section 5.1

K.1 Source of sparsity in single-cell CRISPR screens

Recall in such experiments, each cell receives several perturbations targeting different genome elements. Due to such pooling of a large number of perturbations in a single experiment, most perturbations are present in only a small fraction of cells. Furthermore, gene expression data are measured as RNA molecule counts, and when measured at single-cell resolution, the relatively small number of total RNA molecules measured per cell and the large number of genes result in many genes having zero expression in most cells (Svensson, 2020).

K.2 Parameters and methods implementation in Section 5.1

Parameters used in Section 5.1 We adopt the parameter settings displayed in Table 3. Note that the bolded values of -5 for γ_0 and β_0 are the default parameter values. Instead of testing all combinations of these two parameters, we vary one of them while fixing the other to -5. Furthermore, note that our choices of ρ differ based on whether we are carrying out left- or right-sided tests.

γ_0	β_0	ρ (left-sided)	ρ (right-sided)	r	n
-6	-6	-4	0	0.05	5000
-5	**-5**	-3	0.5	1	
-4	-4	-2	1	10	
-3	-3	-1	1.5		
-2	-2	0	2		

Table 3: Simulation parameter choices.

Methods details in Section 5.1

- The **spaCRT** (Algorithm 2), where $\boldsymbol{X} \mid \boldsymbol{Z}$ is fit based on a logistic regression model and $\boldsymbol{Y} \mid \boldsymbol{Z}$ is fit based on a negative binomial regression model. The size parameter r is estimated by applying the method of moments to the residuals of the Poisson regression of Y on Z (Barry et al., 2021; Barry et al., 2024). This method (called "precomputed" in Table 4) is fast but less accurate than maximum likelihood estimation, but is sufficient for the spaCRT, which does not

require accurate estimation of the size parameter. We use the `uniroot` function in `R` to solve the equation saddlepoint equation. When the solution is not found or the resulting p-value p_{spaCRT} is not in the range $[0, 1]$, we use the p-value based on the GCM test as a backup (see below). We found the failure to solve the saddlepoint equation quite rare, occurring in at most 1.3% of replications across all simulation settings.

- The **dCRT** (Algorithm 1), with the same fitting procedures as the spaCRT and $M = 10{,}000$.
- The **GCM test** (Shah and Peters, 2020), which is based on the asymptotically normal test statistic

$$T_n^{\text{GCM}} \equiv \frac{T_n^{\text{dCRT}}(X, Y, Z)}{\widehat{S}_n}, \ \widehat{S}_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n} R_{in}^2 - \left(\frac{1}{n}\sum_{i=1}^{n} R_{in}\right)^2,$$

  where $T_n^{\text{dCRT}}(X, Y, Z)$ is defined as in (17) and

$$R_{in} \equiv (X_{in} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})).$$

  We use the same fitting procedures for the GCM test as for spaCRT and dCRT.
- The **negative binomial regression score test** (implemented via the `glm.nb()` function in the `MASS` package). This function computes the maximum likelihood estimate of the size parameter iteratively, which is slower than the precomputed approach but more accurate. We choose this approach since the score test relies more heavily on the accuracy of the size parameter estimate.

For negative binomial regression, the estimation of size parameter is a crucial step. GCM, dCRT and spaCRT use the simple method of moments estimator, while the score test uses a more sophisticated estimator that requires iterative computation. The detailed discussion of these dispersion estimation methods and other details of testing methods can be found in Appendix K.2. The comparison of different methods applied is summarized in Table 4. We applied both left- and right-sided variants of each test. All simulations are repeated $10,000$ times for accurate Type-I error estimation for small p-value thresholds.

Test	Dispersion estimation	Resampling required	Normality based
GCM test	Precomputed	No	Yes
Score test	Iterative	No	Yes
dCRT	Precomputed	Yes	No
spaCRT	Precomputed	No	No

Table 4: Summary table for testing methods compared.

K.3 Additional simulation results in Section 5.1

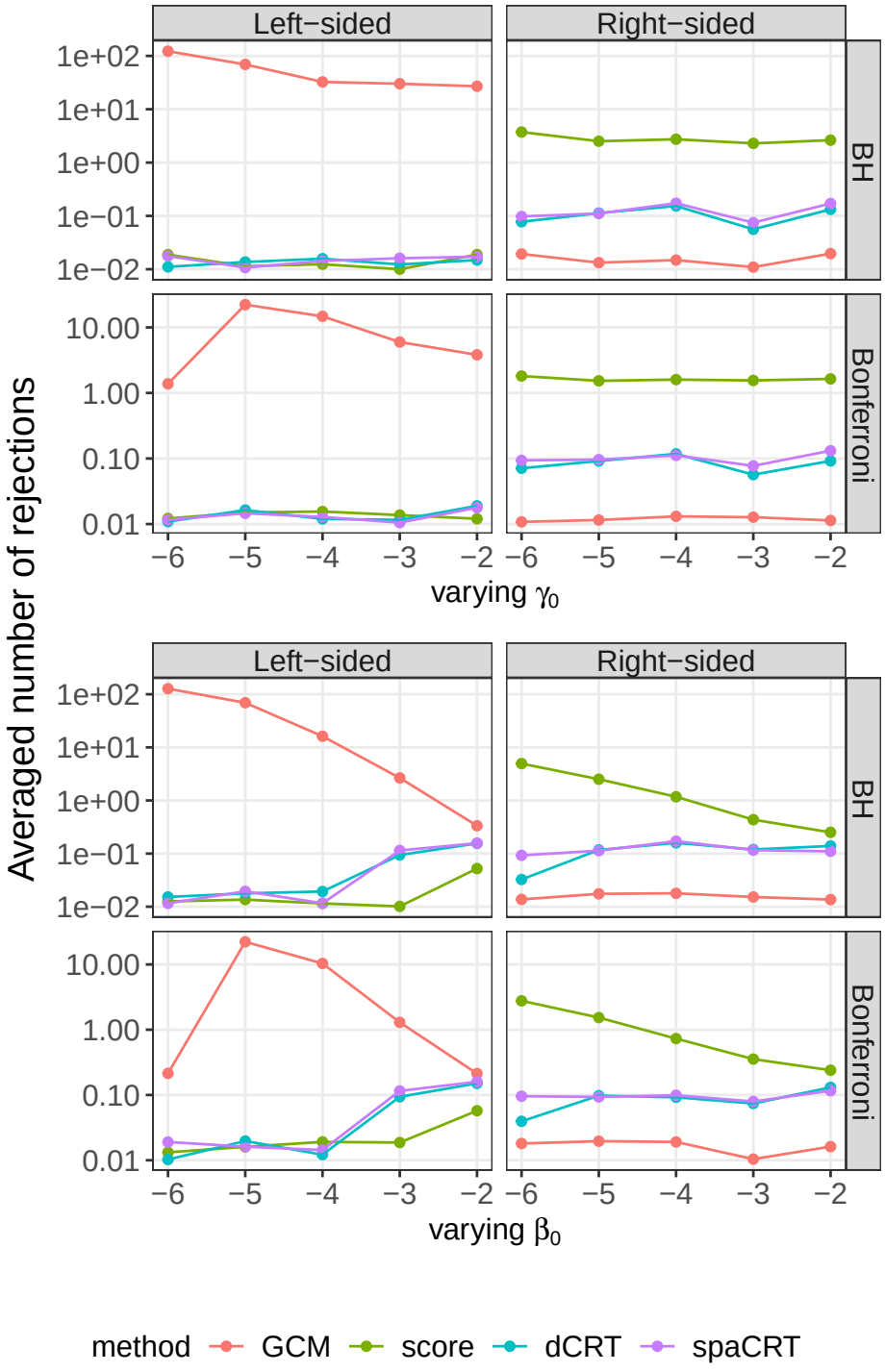

Figure 6: Averaged number of rejections after BH and Bonferroni corrections under the setup $r = 0.05$.

L Additional simulation details in Section 5.2

L.1 Source of sparsity in GWAS with rare phenotypes

GWAS aims to investigate if genetic variation is associated with a phenotype of interest. There can be two sources of sparsity in such analysis. On one hand, the phenotype of interest (Y) can be rare, i.e. only a small fraction of the population has the phenotype. Typical examples include certain rare diseases which have low prevalence in the population. On the other hand, the sparsity can come from the rare genetic variation (X), i.e., only a small fraction of the population has the genetic variation.

L.2 Parameters and methods implementation in Section 5.2

Parameters used in Section 5.2 Recall the logistic regression model:

$$\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X}) \stackrel{d}{=} \text{Ber}(\text{expit}(\gamma_0 + \boldsymbol{X}^\top \beta))$$

where γ_0 is an intercept term, β is a vector of coefficient and g is a smooth function. For the concrete choice of β, we consider

$$\beta = (\underbrace{\eta, \ldots, \eta}_{0.05*p}, \underbrace{-\eta, \ldots, -\eta}_{0.05*p}, \underbrace{0, \ldots, 0}_{0.9*p})^\top$$

where $\eta > 0$ is a signal strength. We vary $\eta \in \{0, 0.25, 0.5, 0.75, 1\}$. For γ_0, we consider $\{-3, -2\}$ for *high* and *low* sparsity settings. For the distribution of $\boldsymbol{X}$, we consider the mHHM (Definition 2) where we set the support of the hidden variable $\boldsymbol{U}_j$, K, to be 10 and that of $\boldsymbol{X}$, M, to be 2. The Markov transition matrix is

$$Q \equiv \begin{pmatrix} \gamma & 1-\gamma & 0 & \ldots & 0 & 0 \\ 0 & \gamma & 1-\gamma & \ldots & 0 & 0 \\ \vdots & & \ddots & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \ldots & \gamma & 1-\gamma \\ 0 & 0 & 0 & \ldots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{K \times K}.$$

We set $\gamma = 0.9$ to create non-trivial correlation between $\boldsymbol{X}_j$. The intial distribution q is a uniform distribution over the support of $\boldsymbol{U}_1$. Besides the transition matrix, we also vary the emission distribution $\boldsymbol{X}_j|\boldsymbol{U}_j$ by considering a beta-piror emission distribution:

$$\mathbb{P}[\boldsymbol{X}_j = 1|\boldsymbol{U}_j = k] \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta). \qquad \text{(beta-emission)}$$

Hyperparameter (α, β) controls the shape of the Beta distribution. We consider two choices: $(\alpha, \beta) = (1, 3)$ and $(\alpha, \beta) = (1, 1)$. The first set of parameters will put more mass towards small values close to 0, which induce high sparsity and thus mimic the rare genetic variation setup. The second choice is a uniform distribution over $[0, 1]$ with no skewness. Thus this induce low sparsity in X and thus mimic the common genetic variation scenario. The simulation setup can be summarized in Table 5.
We will consider regularization parameters λ used in `glmnet` to be $\lambda =$ `lambda.1se` or $\lambda =$ `lambda.min`.

Table 5: Parameters considered in GWAS simulation.

Parmeters $(\alpha, \beta, \gamma_0)$	Sparsity level for (X, Y)
$(1, 3, -2)$	(high, low)
$(1, 3, -3)$	(high, high)
$(1, 1, -2)$	(low, low)
$(1, 1, -3)$	(low, high)

Methods details in Section 5.2

- The **spaCRT** (Algorithm 2), where the parameters of the HMM, the distribution $\boldsymbol{X}$, is fitted based on expectation-maximization (EM) algorithm implemented by `fastPhase` software (Scheet and Stephens, 2006) and `fastPhase` has also been a popular method in the recent variable selection literatures (Sesia, Sabatti, and Candès, 2019). Conditional distribution $\boldsymbol{Y} \mid \boldsymbol{X}_j$ is fitted based on a modified high-dimensional logistic regression with lasso penalization (Tibshirani, 1996). We refer the details of methods implementation to Appendix L.3.
- The **Knockoffs** (Barber and Candès, 2015; Sesia, Sabatti, and Candès, 2019), where the distribution of $\boldsymbol{X}$ is fitted in the same way as in spaCRT and knockoff variables are constructed via backford sampling algorithm proposed in Sesia, Sabatti, and Candès (2019). The test statistic is chosen to be the difference of absolute coefficient values between the variable of interest $\boldsymbol{X}_j$ and its corresponding knockoff variable $\tilde{\boldsymbol{X}}_j$, which are obtained in the high-dimensional logistic regression with lasso penalization for fitting $\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{X}}$, where $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{X}}_1, \ldots, \tilde{\boldsymbol{X}}_d)^\top \in \mathbb{R}^d$.

We also include **dCRT** (Algorithm 1) with resample $M = 5000$ and **GCM test** (Shah and Peters, 2020), with the same fitting procedures as the spaCRT.

L.3 Tower-trick for spaCRT, dCRT and GCM

We devote this section to discussing the computational details of each method and we will discuss a simple yet powerful computational trick to compute the leave-one-out conditional expectations $\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X}_{-j}]$ for dCRT, spaCRT and GCM. We first observe the following identity:

$$\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}_{-j}] = \mathbb{E}[\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]|\boldsymbol{X}_{-j}]$$

where the outer expectation is taken with respect to the measure $\boldsymbol{X}_j|\boldsymbol{X}_{-j}$. If we can estimate the joint distribution of $\boldsymbol{X}$ from data X and one regression estimate for $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$, computing the conditional expectation $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}_{-j}]$ for any $j \in [p]$ is straightforward via the integral evaluation with respect to measure $\boldsymbol{X}_j|\boldsymbol{X}_{-j}$, without any additional regression fit. In other words, we only need one regression fit for $\mathbb{E}[\boldsymbol{Y}|\boldsymbol{X}]$ and one joint distribution fit for $\boldsymbol{X}$. In practice, we consider using the following algorithm:

1. **Estimate distribution $\boldsymbol{X}$:** this can be done by using `fastPhase` (Scheet and Stephens, 2006);

2. **Compute regression estimate $\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X} = x]$:** this can be done by using `glmnet` (Tibshirani, 1996) with the family set to be `binomial`;

3. **Compute $\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X}_{-j}]$ for any $j \in [d]$:** this can be done by computing the following integral:

$$\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X}_{-j} = x_{-j}] = \sum_{x_j \in \{0,1\}} \widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X} = x]\widehat{\mathbb{P}}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = x_{-j}],$$

where $\widehat{\mathbb{P}}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = \cdot]$ is estimated using the `fastPhase` algorithm.

L.4 Additional simulation results in Section 5.2

We present four simulation results with each corresponding to the choice of `lambda.min` or `l ambda.1se` and the sparsity level in X. High sparsity in X with `lambda.1se` has been presented in Figure 3. The other three plots can be found in Figure 7,8 and 9.

We want to point out that the choice of λ seems to affect both the FDR and power of the methods. In particular, comparing Figure 7 and Figure 8, we can find the FDR can be slightly inflated when `lambda.min` is used for dCRT, GCM and spaCRT methods whereas the FDR is well controlled when `lambda.1se` is used. The inflation of false positive rate can be because of the tower trick used for these methods. The intuition is that when `lambda.1se` is used for dCRT, GCM and spaCRT, the estimated regression coefficients are more sparse and the models obtained from `glmnet` is less variable whereas `lambda.min` will lead to more variable models due to the relatively dense model it will produce. On the power side, we can see that the power of dCRT, GCM and spaCRT is slightly improved when `lambda.min` is used. Interestingly, the FDR of Knockoff procedure seems to be more robust to the choice of λ whereas the power is worse when `lambda.min` is used, which is different from the behavior of dCRT, GCM and spaCRT.

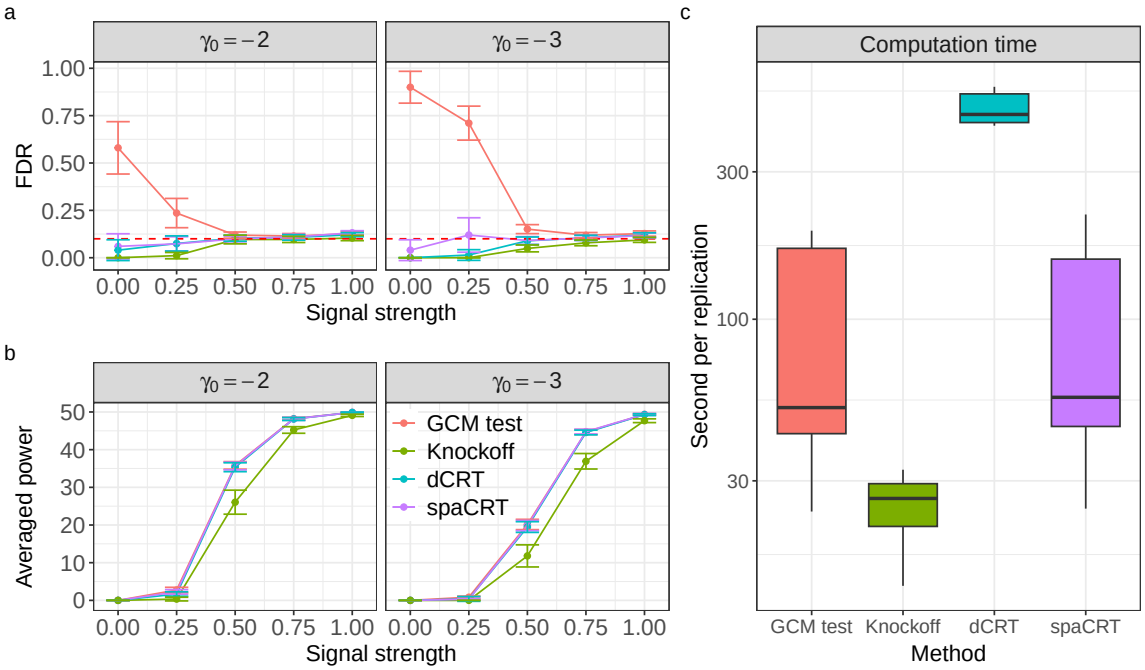

Figure 7: Summary of numerical simulation results for variable selection with low sparsity in data X. All the results are obtained with the regularization parameter $\lambda =$ `lambda.1se`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times consumed by different methods.

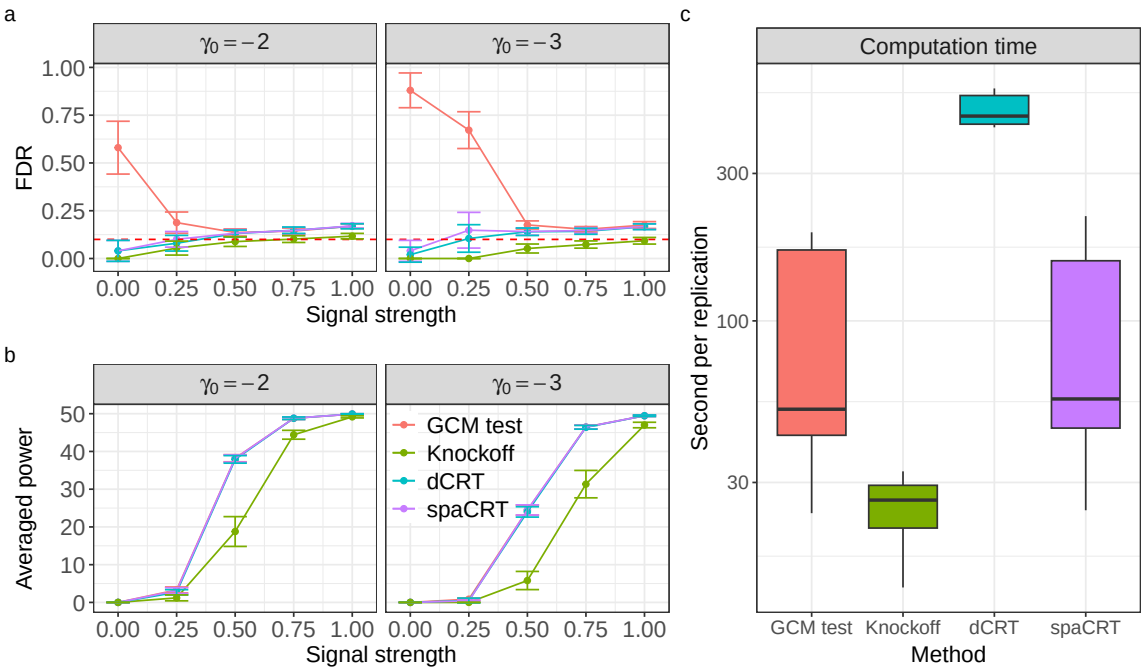

Figure 8: Summary of numerical simulation results for variable selection with low sparsity in data X. All the results are obtained with the regularization parameter $\lambda =$ `lambda.min`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times consumed by different methods.

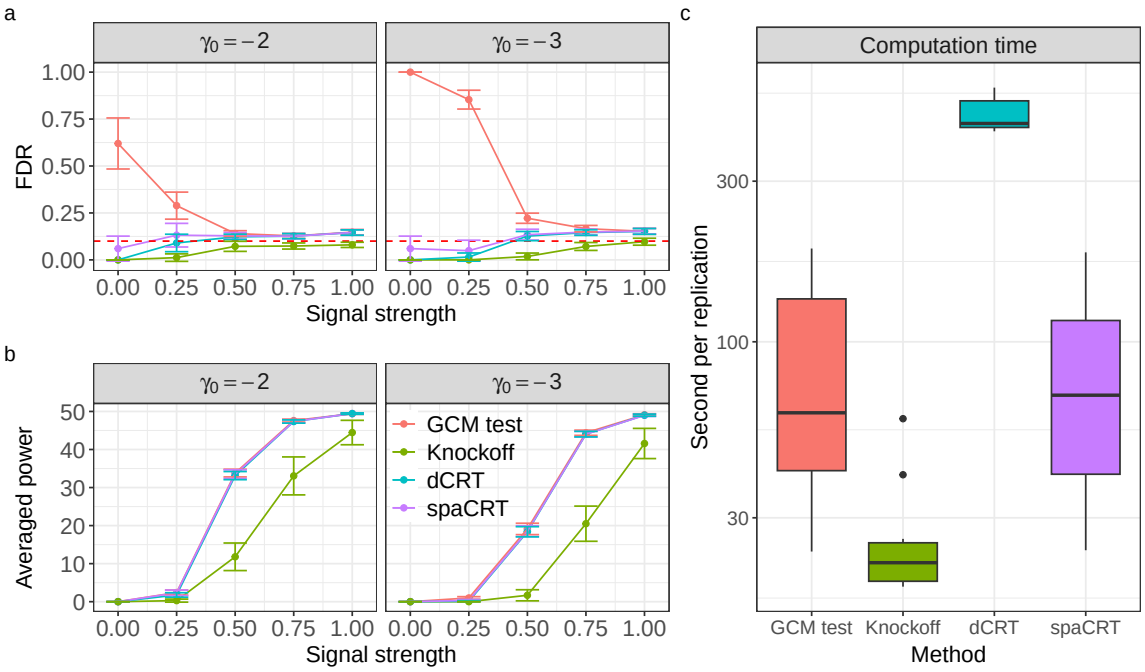

Figure 9: Summary of numerical simulation results for variable selection with high sparsity in data X. All the results are obtained with the regularization parameter $\lambda = \texttt{lambda.min}$. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times consumed by different methods.

M Application to nonparametric regression

M.1 Theoretical results on kernel ridge regression

In this section, we study the validity of spaCRT when the conditional distribution $\boldsymbol{Y}|\boldsymbol{Z}$ is modeled using *kernel ridge regression* (KRR), a representative of nonparametric machine learning methods. Throughout this section, we will assume we are under the classical low-dimensional setup so that we can simplify the subscript $(Z_{in}, X_{in}, Y_{in}) = (Z_i, X_i, Y_i)$ and $\theta_n = \theta$.

Suppose the conditional expectations $\mu_{n,y} \in \mathcal{H}$ for some RKHS $(\mathcal{H}, \|\cdot\|_{\mathcal{H}})$ with reproducing kernel $k \in \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. Let $K \in \mathbb{R}^{n\times n}$ have ijth entry $K_{ij} = k(Z_i, Z_j)/n$ and denote the eigenvalues of K by $\widehat{\kappa}_1 \geq \widehat{\kappa}_2 \geq \cdots \geq \widehat{\kappa}_n \geq 0$. We will assume that kernel function k admits an eigen-expansion of the form

$$k(z, z') = \sum_{j=1}^{\infty} \kappa_j e_j(z) e_j(z') \tag{92}$$

with orthonormal eigenfunctions $\{e_j\}_{j=1}^{\infty}$, so $\mathbb{E}[e_j e_k] = \mathbb{1}(k = j)$, and summable eigenvalues $\kappa_1 \geq \kappa_2 \geq \cdots \geq 0$. Such expansion can be guaranteed by Mercer's theorem if mild conditions are satisfied. For a sequence of regularization parameter λ_n, we consider the following estimator:

$$\widehat{\mu}_y \equiv \arg\min_{\mu_y \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_y(Z_i))^2 + \lambda_n \|\mu_y\|_{\mathcal{H}}^2 \right\}. \tag{93}$$

We consider selecting λ_n in the following data-dependent way:

$$\lambda_n = \arg\min_{\lambda > 0} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{\kappa}_i^2}{(\widehat{\kappa}_i + \lambda)^2} + \lambda \right\} \tag{94}$$

We want to emphasize that the way we select the tuning parameter is mainly for the ease of theoretical analysis and similar data-dependent hyperparameter selection has been adopted in previous work Niu et al. (2024) and Shah and Peters (2020). As for the estimator $\widehat{\theta}$ for θ, we consider using the maximum likelihood estimator $\widehat{\theta}$. With the estimators $\widehat{\mu}_y(Z_i)$ and $\widehat{\theta}_{n,x}(Z_i) = Z_i^\top \widehat{\theta}$, the spaCRT can be applied with Algorithm 2. Now we state our main results on the validity guarantee of spaCRT.

Theorem 9. *Consider $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z}$. Suppose assumptions 3-4 hold. Then if the following conditions hold:*

$$\text{support of } \boldsymbol{Y} \text{ is compact, i.e., there exists } S, \mathbb{P}[\boldsymbol{Y} \in [-S, S]] = 1; \tag{95}$$

$$\|\widehat{\theta} - \theta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}); \tag{96}$$

$$\|\widehat{\mu}_y\|_\infty = O_{\mathbb{P}}(1); \tag{97}$$

$$\sum_{j=1}^{\infty} \kappa_j < \infty, \tag{98}$$

then the conclusion in Theorem 2 hold and $\lim_{n\to\infty} \mathbb{E}[\phi_{n,\alpha}^{\text{spaCRT}}] = \alpha$.

Conditions in Theorem 9 are mild conditions. Condition (97) can be easily verified when linear kernel, i.e. linear ridge regression, is considered. For general choice of kernel, condition (97) can hold under extra conditions on the kernel function k.

M.2 Simulation on unbalanced nonparametric classification

N Additional figures and tables for real data anlaysis

We present relevant information about Gasperinit dataset in Section N.1. In Section N.2, we show a table including the number of rejections when applying Bonferroni or BH method to the pairs invlovling the non-targeting perturbations (thus under the null). The total number of hypotheses is 153000. In Section N.3, we present additional figures for real data analysis including QQ-plots faceting across different effective sample size (Figure 10) and QQ-plots faceting across different dispersion parameters (Figure 11).

N.1 Overview of the data

The Gasperini data contain expression measurements on 13,135 genes and CRISPR perturbations targeting 6,105 regulatory elements in n = 207,324 cells. They also contain CRISPR perturbations intended as negative and positive controls. In particular, the data contain 51 non-targeting CRISPR perturbations, which do not target any regulatory element and therefore should have no effect on the expressions of any genes. Furthermore, the data contain 754 CRISPR perturbations targeting genes, rather than regulatory elements. These serve as positive controls, because they are known a priori to have effects on the expressions of the genes they target. Finally, the data contain measurements on six covariates, including four count-based covariates related to library size, one binary covariate indicating the experimental batch, and one continuous covariate indicating the proportion of reads mapping to mitochondrial genes in each cell.

N.2 Additional tables for the real data analysis

Table 6: Number of rejections for negative control pairs on the Gasperini data.

	Number of rejections			
	Left-sided test		Right-sided test	
Method	Bonferroni	BH	Bonferroni	BH
GCM test	22	128	0	0
Score test	1	1	15	29
spaCRT	1	1	0	0
dCRT	1	4	0	0

N.3 Additional figures for the real data analysis

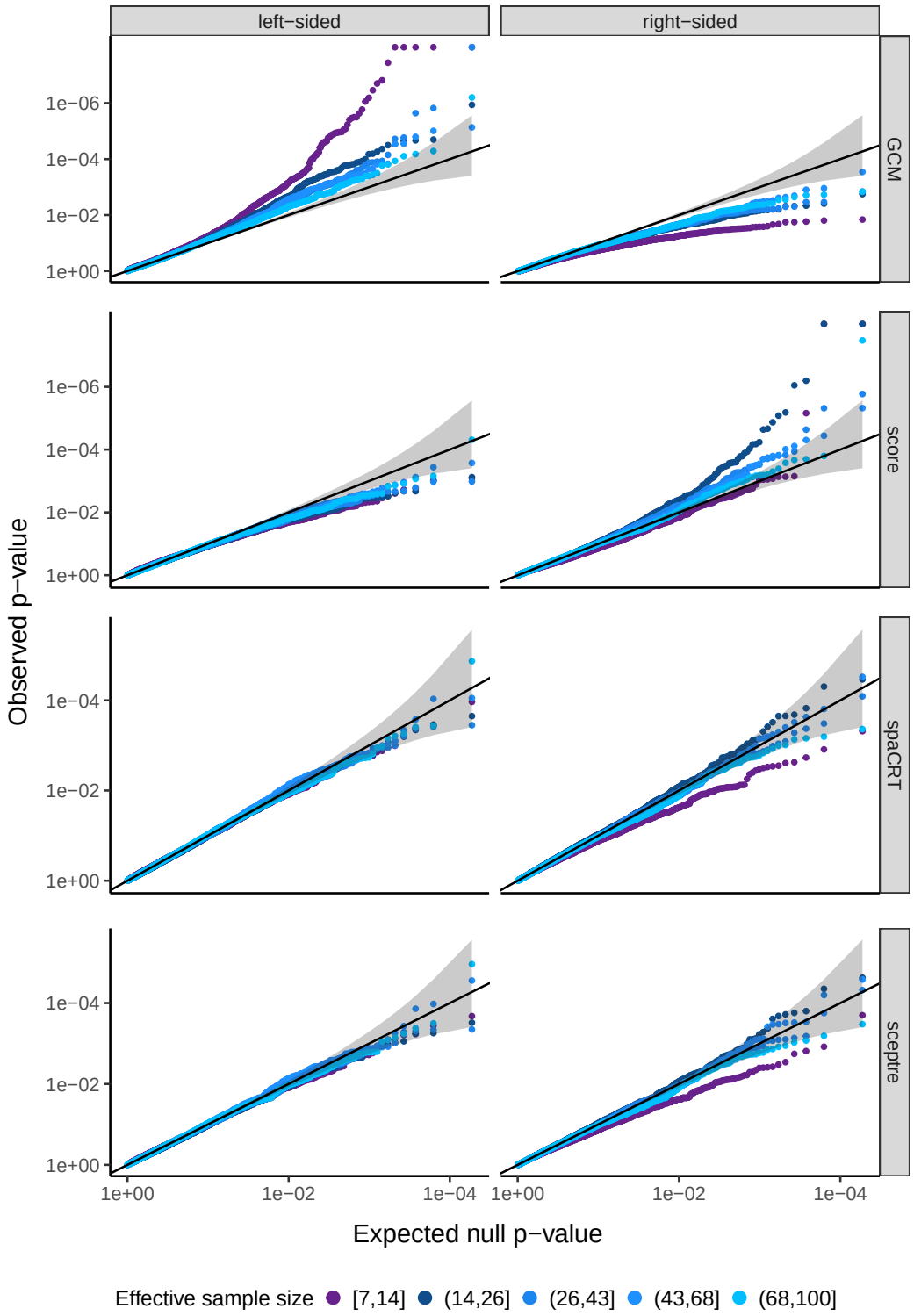

Figure 10: QQ-plots for the p-values of right-sided test from different methods under low effective sample size.

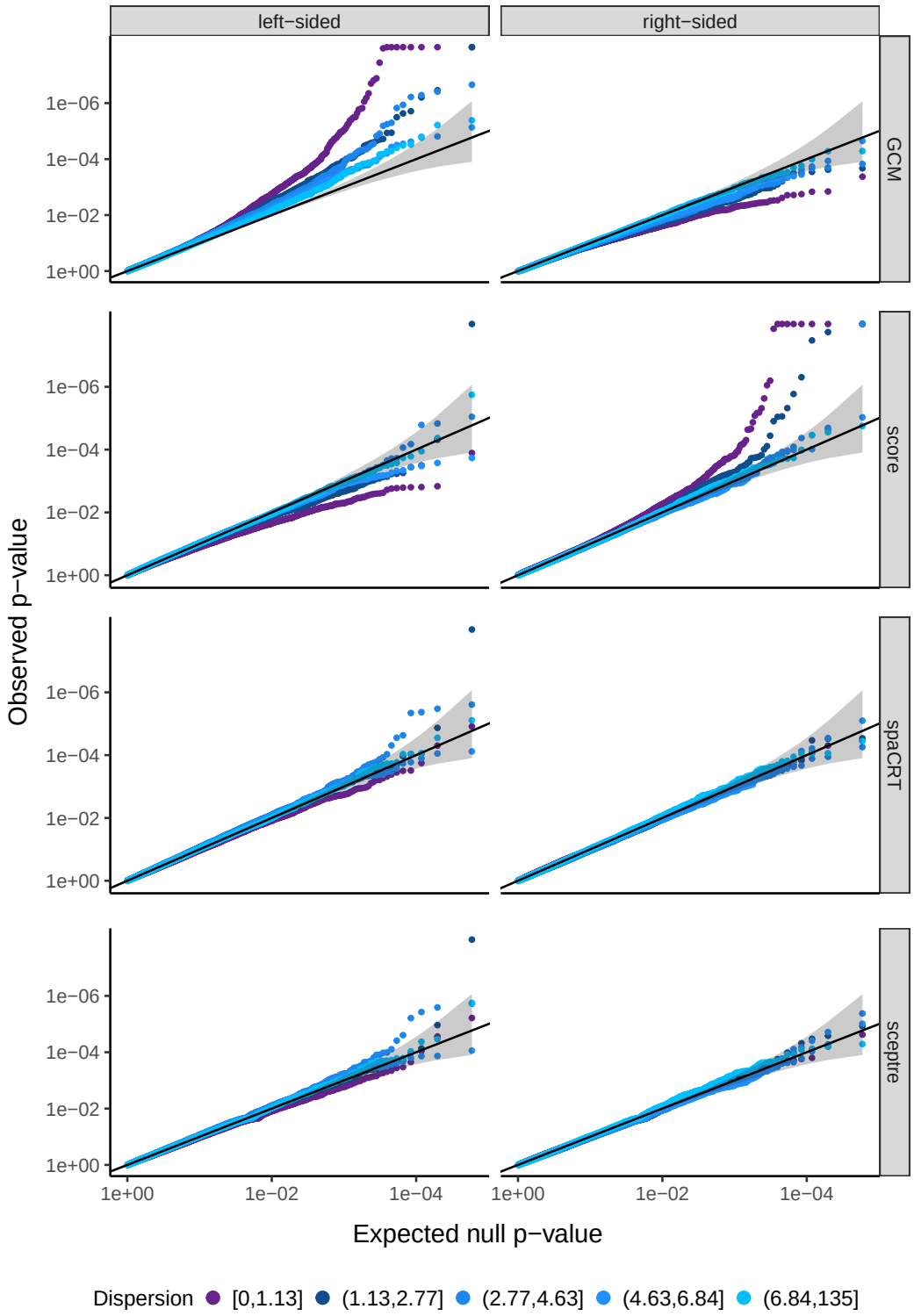

Figure 11: QQ-plots for the p-values of left-sided test from different methods stratified by dispersion parameter.