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

Ziang Niu, Jyotishka Ray Choudhury, Eugene Katsevich

June 13, 2025

Abstract

Saddlepoint approximations (SPAs) for resampling-based procedures offer statistically accurate and computationally efficient inference, which is particularly critical in the analysis of large-scale, high-multiplicity data. Despite being introduced 70 years ago, SPAs for resampling-based procedures lack rigorous justification and have been underutilized in modern applications. We establish a theoretical foundation for the SPA in this context by developing a general result on its approximation accuracy for conditional tail probabilities of averages of conditionally independent summands. This result both justifies existing SPAs for classical procedures like the sign-flipping test and enables new SPAs for modern resampling methods, including those using black-box machine learning. Capitalizing on this result, we introduce the saddlepoint approximation-based conditional randomization test (spaCRT), a resampling-free conditional independence test that is both statistically accurate and computationally efficient. The method is especially well-suited for sparse, large-scale datasets such as single-cell CRISPR screens and genome-wide association studies involving rare diseases. We prove the validity of the spaCRT when paired with modern regression tools such as lasso and kernel ridge regression. Extensive analyses of simulated and real data show that the spaCRT controls Type-I error, achieves high power, and outperforms existing asymptotic and resampling-based alternatives.

Keywords: conditional independence testing, conditional randomization test, resampling, saddlepoint approximation, single-cell CRISPR screens.

1 Introduction

1.1 Multiplicity and sparsity in modern hypothesis testing

Modern data collection technologies have enabled the generation of datasets with large sample sizes and many variables of interest for statistical testing. In many cases, these datasets are also highly sparse, in the sense that a large proportion of the observations are zeros. Such data arise in various domains, such as risk variable selection for sparse but high-impact insurance claims (Wang, Ma, and Wang, 2015), detection via genome-wide association studies (GWAS) of rare genetic variants involved in rare diseases (Dey et al., 2017; Zhao et al., 2020), CRISPR screens with single-cell readouts (Dixit et al.,

2016; Adamson et al., 2016), and sparse rating comparison between underrepresented subgroups in recommender systems (Yao and Huang, 2017).

The combination of high multiplicity and sparsity in these applications poses significant challenges for statistical inference. Multiplicity leads to stringent significance thresholds, focusing attention far into the tails of null distributions. When applied to highly sparse data, the normal approximations underlying most statistical tests can become unreliable in these tail areas, as has been demonstrated empirically in both GWAS (Dey et al., 2017) and single-cell CRISPR screens (Barry et al., 2024). While resampling-based procedures can provide more accurate null distributions, their computational cost can be prohibitive for high-multiplicity problems. In addition to having to repeat the resampling for each hypothesis, stringent significance thresholds necessitate a large number of resamples per hypothesis to achieve small enough p-values. Therefore, practitioners face a difficult trade-off between statistical accuracy and computational efficiency.

1.2 Existing approaches

The tension between statistical accuracy and computational efficiency is a common theme across many statistical applications. Several strands of work have been developed to address this challenge.

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, 2024; Fischer, Barry, and Ramdas, 2024). Such methods can substantially reduce the number of resamples needed, and are applicable to arbitrary resampling schemes and test statistics. However, they typically require at least hundreds of resamples to achieve good power, which can be computationally expensive in large-scale settings; more aggressive reductions in the number of resamples tend to result in a loss of power. Furthermore, adaptive resampling schemes do not provide accurate p-value approximations (by design), which can limit the interpretability of their outputs for practitioners accustomed to visualizations like Manhattan plots (standard in GWAS analysis), volcano plots, or QQ plots. Another approach to reducing the number of resamples is to fit a parametric curve to the resampling distribution based on a small number of resamples. This heuristic approach has been employed in several statistical applications (Ge et al., 2012; Winkler et al., 2016; Barry et al., 2021), but lacks theoretical justification.

A different class of methods addressing this dilemma leverages refined asymptotic approximation techniques, which provide accurate, closed-form p-value approximations while circumventing resampling. The saddlepoint approximation (SPA; Daniels, 1954; Lugannani and Rice, 1980) is a classical method known for its highly accurate approximations of tail probabilities. A key advantage of the SPA is its relative error guarantee (Kolassa, 2006; Butler, 2007), especially valuable in high-multiplicity settings where the accurate approximation of small p-values is crucial. The SPA has been proposed as an approximation for the resampling distributions used in classical procedures such as permutation tests (Robinson, 1982) and the bootstrap (Davison and Hinkley, 1988). Despite its promise, the application of the SPA in this context

brings theoretical challenges, including accommodating the conditioning inherent in resampling-based procedures and the non-smoothness of their resampling distributions. As a result, there currently is no rigorous justification for the SPA in the context of resampling-based hypothesis testing. Furthermore, the SPA is viewed as a rather classical technique with limited application in modern statistical practice, especially in contexts with high-dimensional and nonparametric nuisance parameters.

1.3 Our contributions

To facilitate theoretically grounded, statistically accurate, and computationally efficient testing in sparse, high-multiplicity settings, we pursue the SPA approach. To overcome the aforementioned challenges with this approach, we first lay a general theoretical foundation for the SPA in the context of resampling-based hypothesis testing. Then, we build on this foundation to propose an SPA-based procedure for a prototypical modern hypothesis testing problem: conditional independence (CI) testing. CI testing is a fundamental building block in variable selection (Candès et al., 2018), causal inference (Pearl, 2009), and graphical models (Lauritzen, 1996; Koller and Friedman, 2009), and has been applied in a wide range of domains (Magwene and Kim, 2004; Sesia, Sabatti, and Candès, 2019; Barry et al., 2021; Sekulovski et al., 2024). In more detail, the following are our two central contributions:

1. We provide a theoretical justification for SPA in the context of resampling-based hypothesis testing, accounting rigorously for conditioning. This justifies the SPA for classical resampling-based procedures, like the sign-flipping test (Daniels, 1955), 70 years after these approximations were first proposed. Our results also loosen the assumptions of the SPA, not requiring continuity or lattice assumptions, establishing a new result even for the classical (unconditional) SPA.

2. We apply the SPA result to the resampling-based distilled conditional randomization test (dCRT; Liu et al., 2022) 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 at github.com/Katsevich-Lab/spacrt.

Through simulation studies modeled on 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 (Gasperini et al., 2019) with roughly 250,000 observations, whose original analysis included testing 84,595 hypotheses. Figure 1 previews the results (for details, see Section 6). The resampling-based dCRT gives reliable Type-I error control but would have required 2.7 CPU-years to test all hypotheses. The normality-based GCM test (Shah and Peters, 2020), using essentially the same test statistic, gives severely inflated p-values. The spaCRT performs as well as the dCRT statistically while matching the speed of the GCM test, requiring just 3.9 CPU-days and thereby providing an approximately 250-fold speedup.

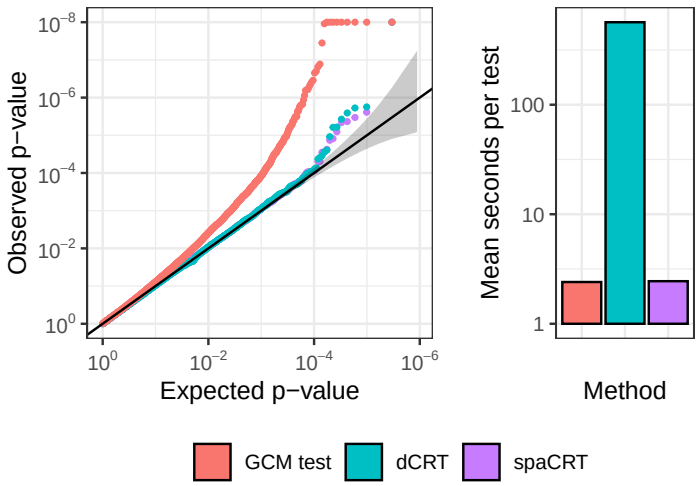

Figure 1: Comparing the Type-I error control and computation times of the proposed spaCRT with other methods on the Gasperini et al. (2019) data. Left: QQ-plot of the p-values under the null hypothesis. Right: Mean computation times per hypothesis, in seconds.

Paper outline. We introduce our general theorem on the accuracy of the SPA for conditional tail probabilities in Section 2. We then present the spaCRT methodology in Section 3. We provide theoretical results for this methodology in Section 4, simulation results in Section 5 and a real data application in Section 6, respectively. We conclude with a discussion in Section 7.

2 The conditional SPA

Consider a resampling-based hypothesis test using the statistic $T_n \equiv \frac{1}{n}\sum_{i=1}^n T_{in}$ computed based on a dataset $\mathcal{D}_n$. The resampling distribution of this test is $\widetilde{T}_n \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{T}_{in}$, for some resampling mechanism generating $\widetilde{T}_{in}$ conditionally on the data $\mathcal{D}_n$. Then the p-value of this test can be defined as

$$p_n \equiv \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{T}_{in} \geq \frac{1}{n}\sum_{i=1}^n T_{in} \;\middle|\; \mathcal{D}_n\right]. \tag{1}$$

To abstract the statistical problem, 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:

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

matching the definition (1) by setting $W_{in} \equiv \widetilde{T}_{in}$, $w_n \equiv T_n$, and $\mathcal{F}_n \equiv \sigma(\mathcal{D}_n)$. These assumptions on W_{in} encompass diverse resampling-based procedures, such as the bootstrap (Efron, 1979) and the sign-flipping test (Daniels, 1955), but excluding permutation tests, which we leave for future work. In this section, we state our main results

on the approximation accuracy of a conditional Lugannani-Rice (Lugannani and Rice, 1980) SPA formula to approximate the tail probability (2) (Section 2.1). We then discuss how this result relates to existing results (Section 2.2). In this section and throughout, we use standard asymptotic notations $O_{\mathbb{P}}, o_{\mathbb{P}}$, and $\Omega_{\mathbb{P}}$, whose formal definitions can be found in Appendix A.1.

2.1 Accuracy of the conditional saddlepoint approximation

We first extend the definitions of sub-exponential and compactly supported random variables to the conditional setting, obtaining the *conditionally sub-exponential* (CSE) and *conditionally compactly supported* (CCS) conditions.

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,*

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

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,*

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

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

$$K_{in}(s) \equiv \log \mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] \quad \text{and} \quad K_n(s) \equiv \frac{1}{n}\sum_{i=1}^n K_{in}(s). \tag{3}$$

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

$$K_n'(s) = w_n. \tag{4}$$

We denote the solution to saddlepoint equation (4) as $\hat{s}_n$, whose existence and uniqueness Theorem 1 below guarantees. With the solution $\hat{s}_n$ in hand, we define the saddlepoint approximation as follows. If we define

$$\lambda_n \equiv \hat{s}_n \sqrt{n K_n''(\hat{s}_n)}; \quad r_n \equiv \begin{cases} \operatorname{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_n w_n - K_n(\hat{s}_n))} & \text{if } \hat{s}_n w_n - K_n(\hat{s}_n) \geq 0; \\ \operatorname{sgn}(\hat{s}_n) & \text{otherwise,} \end{cases} \tag{5}$$

then the saddlepoint approximation is given by

$$\widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^n W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \tag{6}$$

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 (6), so that $\widehat{\mathbb{P}}_{\mathrm{LR}} \equiv 1/2$. We also use the convention $0/0 \equiv 1$. The approximation $\widehat{\mathbb{P}}_{\mathrm{LR}}$ (6) is a direct generalization of the classical Lugannani-Rice (LR) formula (Lugannani and Rice, 1980) to the conditional setting. The theorem below establishes the accuracy of this approximation.

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*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[W_{in}^2 \mid \mathcal{F}_n] = \Omega_{\mathbb{P}}(1). \tag{7}$$

Let $w_n \in \mathcal{F}_n$ be a sequence with $w_n \xrightarrow{\mathbb{P}} 0$. Then, there exists $\varepsilon > 0$ such that the saddlepoint equation (4) 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 $\hat{s}_n$ and definition of ε can be found in Appendix C.1. Furthermore, the SPA $\widehat{\mathbb{P}}_{\mathrm{LR}}$ to the conditional tail probability (2) defined by equations (5) and (6) has vanishing relative error:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n}W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \tag{8}$$

We make several remarks on this result.

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 (8) does not come with a rate, unlike most existing results. Nevertheless, we find it useful to have a single result encompassing a broad range of settings, in the same way that it is useful to have the central limit theorem in addition to the Berry-Esseen theorem. 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 Assumption 1 (sub-exponential summands) and condition (7) (non-degenerate variance). These are standard and easy to understand and verify.

Remark 2 (Justification of application to resampling-based tests)**.** The result (8) 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 the CRT, result (8) provides **the first rigorous justification of an SPA for the sign-flipping test** of the kind originally proposed by Daniels (1955), which is summarized as Theorem 5 in Appendix C.3.

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 (4) 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 (8).

Remark 4 (Assumptions on the threshold w_n)**.** The role of the assumption $w_n = o_{\mathbb{P}}(1)$ is to guarantee the existence of a solution the saddlepoint equation beyond the case of compact support (Daniels, 1954). The assumption may appear restrictive at first glance, but since the rate of convergence of w_n towards zero can be arbitrary, it accommodates at least two statistically interesting regimes: (a) Moderate deviation regime: $w_n = O_{\mathbb{P}}(n^{-\alpha}), \alpha \in (0, 1/2)$ and (b) CLT regime: $w_n = O_{\mathbb{P}}(1/\sqrt{n})$.

2.2 Connection to existing unconditional results

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.*

$$\mathbb{P}[|W_{in}| \geq t] \leq \theta \exp(-\beta t) \quad \textit{for all } t > 0. \tag{9}$$

Furthermore, suppose that

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[W_{in}^2] > 0. \tag{10}$$

Given a sequence of cutoffs $w_n \to 0$ as $n \to \infty$, there is an $\varepsilon > 0$ such that the saddlepoint equation (4) has a unique solution $\hat{s}_n$ on $[-\varepsilon/2, \varepsilon/2]$ for all sufficiently large n. Furthermore, the unconditional SPA $\widehat{\mathbb{P}}_{\mathrm{LR}}$ has vanishing relative error

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right](1 + o(1)). \tag{11}$$

Compared to the classical result of Lugannani and Rice (1980), the significance of Theorem 1 and Corollary 1 is the generality and transparency of the assumptions, as discussed in Remark 1. We provide another connection to the classical result of Robinson (1982) in Appendix C.2.

3 spaCRT: A resampling-free approximation to dCRT

The conditional SPA established in Section 2 is more versatile than it appears. In this section, we demonstrate how it can be applied to a resampling-based conditional independence (CI) test, the dCRT (Liu et al., 2022), which accounts for the presence of a potentially high-dimensional covariate vector via black-box machine learning methods. We define the CI testing problem and review the dCRT in Section 3.1. We then introduce the spaCRT, its resampling-free approximation, in Section 3.2.

3.1 Background: CI testing and the dCRT

Consider a predictor variable $\boldsymbol{X}$, an outcome variable $\boldsymbol{Y}$, and a covariate vector $\boldsymbol{Z} \in \mathbb{R}^d$. Given a joint distribution $(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}) \sim \mathcal{L}$, the CI null hypothesis is that the outcome is independent of the predictor, given the covariates:

$$H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}. \tag{12}$$

Suppose that independent and identically distributed observations $(X_{in}, Y_{in}, Z_{in}) \overset{\text{i.i.d.}}{\sim} \mathcal{L}_n$ are collected for $i = 1, \ldots, n$, where the law $\mathcal{L}_n$ is allowed to vary with n. Denoting these observations collectively as $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^n$, $Z \in \mathbb{R}^{n \times d}$, the CI testing problem is to test the null hypothesis (12) based on the data (X, Y, Z).

The dCRT procedure (Liu et al., 2022) is a CI testing procedure, which was initially proposed in the context of *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, reviewed in Appendix D.1). 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

$$\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z}) = f(\boldsymbol{X} \mid \theta = \theta_{n,x}(\boldsymbol{Z})), \quad \text{where} \quad f(x|\theta) = \exp(\theta x - A(\theta))h(x) \tag{13}$$

is an exponential family with natural parameter θ, natural parameter space $\mathbb{R}$, log-partition function A and base measure h. Since we allow $\theta_{n,x}(\boldsymbol{Z})$ to be arbitrary, the model (13) is very flexible. 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, including nonparametric or black-box procedures. Setting $\widehat{\mu}_{n,x}(\boldsymbol{Z}) \equiv A'(\widehat{\theta}_{n,x}(\boldsymbol{Z}))$, we arrive at the test statistic

$$T_n^{\text{dCRT}}(X, Y, Z) = \frac{1}{n} \sum_{i=1}^{n} (X_{in} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})). \tag{14}$$

The dCRT is obtained by comparing $T_n^{\text{dCRT}}(X, Y, Z)$ 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 $T_n^{\text{dCRT}}(X, Y, Z)$ as in (14);

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

$$T_n^{\text{dCRT}}(\widetilde{X}^{(m)}, X, Y, Z) \equiv \frac{1}{n} \sum_{i=1}^{n} (\widetilde{X}_{in}^{(m)} - \widehat{\mu}_{n,x}(Z_{in}))(Y_{in} - \widehat{\mu}_{n,y}(Z_{in})); \tag{15}$$

end

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

As a resampling-based procedure, the dCRT is well-equipped to handle sparse data of the kinds that motivate this work. By contrast, normality-based CI tests like the GCM test (Shah and Peters, 2020) can have poor finite-sample performance in the presence of data sparsity, due to the slower convergence of the central limit theorem (recall Figure 1); see Theorem 7 in Appendix D.2. Despite its appeal, the dCRT can be computationally expensive for large-scale problems due to its resampling. This motivates us to develop a resampling-free approximation to this procedure.

3.2 The spaCRT

To accelerate the dCRT, we propose the spaCRT, a completely resampling-free procedure with nearly identical statistical performance. If we consider the limit of the dCRT p-value as the number of resamples M grows indefinitely, we obtain

$$p_{\text{dCRT}} \equiv \mathbb{P}\left[T_n^{\text{dCRT}}(\widetilde{X}, X, Y, Z) \geq T_n^{\text{dCRT}}(X, Y, Z) \mid X, Y, Z\right]. \tag{16}$$

We approximate this conditional tail probability via the SPA. Note that the resampled test statistic (15) 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 $\mathcal{F}_n \equiv \sigma(X, Y, Z)$. Thus, the dCRT p-value (16) fits the form (2), so we may derive the conditional LR approximation (6) for the dCRT p-value and plug in expressions for the conditional CGF $K_{in}(s)$ and its derivatives under the exponential family model (13) to obtain the spaCRT procedure (Algorithm 2).

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

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

$$\frac{1}{n}\sum_{i=1}^{n} a_{in}\left(A'(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}\hat{s}_n) - A'(\widehat{\theta}_{n,x}(Z_{in}))\right) = T_n^{\text{dCRT}}(X, Y, Z); \tag{17}$$

4 Compute $\lambda_n = \sqrt{n}\hat{s}_n\sqrt{\frac{1}{n}\sum_{i=1}^{n} a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}\hat{s}_n)}$ and $r_n = \text{sgn}(\hat{s}_n)R_n$, where

5 $$R_n = \sqrt{2n\hat{s}_n T_n^{\text{dCRT}} - 2\sum_{i=1}^{n}(A(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}\hat{s}_n) - A(\widehat{\theta}_{n,x}(Z_{in})) - a_{in}\hat{s}_n A'(\widehat{\theta}_{n,x}(Z_{in})))}$$

if the quantity in square root is non-negative; otherwise $R_n = 1$;

Output: spaCRT p-value

$$p_{\text{spaCRT}} \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \tag{18}$$

The spaCRT 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 solving the saddlepoint equation (17). The latter is a one-dimensional root-finding problem that can be solved efficiently. Indeed, since the saddlepoint equations (4) and (17) are based on the increasing function $K_n'(s)$, binary search algorithms can be employed to find the root in logarithmic time.

To exemplify the spaCRT procedure, consider the case where $\boldsymbol{X}$ is binary, as it is in single-cell CRISPR screens, one of our motivating applications.

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

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

where $\mathrm{expit}(x) \equiv 1/(1 + \exp(-x))$. 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 (18).

4 Theoretical guarantees for the spaCRT

In this section, we provide a broad set of conditions under which the spaCRT approximates the dCRT well and controls Type-I error, demonstrating that the spaCRT combines computational speed with statistical accuracy. We present general results in Section 4.1 and explore special cases in Section 4.2.

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:*

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

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

Suppose further that the following conditions hold:

$$|\widehat{\theta}_{n,x}(Z_{in})| < \infty, |\widehat{\mu}_{n,y}(Z_{in})| < \infty \text{ for any } i, n \text{ almost surely;} \tag{19}$$

$$\frac{1}{n}\sum_{i=1}^{n}(Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 A''(\widehat{\theta}_{n,x}(Z_{in})) = \Omega_{\mathbb{P}}(1); \tag{20}$$

$$T_n^{\mathrm{dCRT}}(X, Y, Z) \xrightarrow{\mathbb{P}} 0. \tag{21}$$

Then, the saddlepoint equation (17) 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:

$$p_{\mathrm{dCRT}} = p_{\mathrm{spaCRT}} \cdot (1 + o_{\mathbb{P}}(1)). \tag{22}$$

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

Remark 5 (Comments on conditions)**.** Conditions (CSE) and (CCS) are tail conditions corresponding to Assumptions 1 and 2, respectively. Conditions (19) and (20) are regularity conditions. The role of the condition (21) 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, this condition holds under the null hypothesis and contiguous local alternatives.

Remark 6 (Relative error guarantee)**.** The relative error guarantee in conclusion (22) is a strong result and is a direct consequence of result (8). 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 $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\mathrm{dCRT}} \le \alpha] \le \alpha$, then $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\mathrm{spaCRT}} \le \alpha] \le \alpha$.*

Corollary 2 states the asymptotic validity of spaCRT given the asymptotic validity of dCRT. Under the model-X assumption that $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$ is known exactly, the p-value produced by dCRT procedure is exact, i.e., $\mathbb{P}_{H_0}[p_{\mathrm{dCRT}} \le \alpha] \le \alpha$. Therefore, the spaCRT procedure has asymptotic Type-I error control under the assumptions of Theorem 2. We now discuss the asymptotic validity of spaCRT when considering the general in-sample fit of $\mathcal{L}_n(\boldsymbol{X} \mid \boldsymbol{Z})$.

Remark 7 (Double robustness and asymptotic validity of spaCRT)**.** With in-sample fit $\widehat{\mathcal{L}}_n(\boldsymbol{X} \mid \boldsymbol{Z})$, Niu et al. (2024) showed that the dCRT is asymptotically equivalent to the GCM test (Shah and Peters, 2020) under mild conditions. The GCM test has desirable double robustness properties, which Corollary 2 implies that spaCRT inherits. Shah and Peters (2020) and Smucler, Rotnitzky, and Robins (2019) showed that the GCM test is *rate and model doubly robust*, meaning that it controls Type-I error asymptotically if (1) the product of the rates at which $\mu_{n,x}(\boldsymbol{Z})$ and $\mu_{n,y}(\boldsymbol{Z})$ are learned is faster than $n^{-1/2}$, or (2) one of the two regression functions is estimated consistently while the other estimator converges sufficiently fast—even if its underlying model is misspecified (see details in Appendix D.1). **We conclude that the spaCRT enjoys both of these double robustness properties as well, yielding a wide range of conditions under which it controls Type-I error asymptotically.**

4.2 Case studies with modern regression techniques

We dedicate this section to verifying the approximation accuracy and Type-I error control of the spaCRT in special cases, including those where $\boldsymbol{Y}|\boldsymbol{Z}$ is estimated using modern regression techniques. We consider low- and high-dimensional GLMs in the main text and nonparametric kernel ridge regression in Appendix E.3.

Throughout this section, we will consider the following GLMs for the data:

$$\boldsymbol{X}|\boldsymbol{Z} \sim \text{Ber}(\boldsymbol{X} \mid \text{logit}(\mu) = \boldsymbol{Z}^\top \gamma_n) \quad \text{and} \quad \boldsymbol{Y} \mid \boldsymbol{Z} \sim f(\boldsymbol{Y} \mid \theta = \boldsymbol{Z}^\top \beta_n),$$

for some exponential family $f(y|\theta)$ with natural parameter θ and log-partition function A_y, recalling equation (13). The choice of logistic model for $\boldsymbol{X}|\boldsymbol{Z}$ is mainly for theoretical convenience. In fact, the spaCRT can be easily integrated with diverse models beyond just logistic regression under binary $\boldsymbol{X}$, such as hidden Markov models, which we employ in our simulations (Section 5.2).

We require the following two assumptions:

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* $C_Z \in (0, \infty)$.

Low-dimensional generalized linear regression. Suppose we are under the classical low-dimensional setup so that we write $(X_{in}, Y_{in}, Z_{in}) = (X_i, Y_i, Z_i), \gamma_n = \gamma$ and $\beta_n = \beta$.

Theorem 3. *Suppose* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$*, and that Assumptions 3-4 hold. If the maximum likelihood estimates (MLEs)* $\widehat{\beta}, \widehat{\gamma}$ *satisfy*

$$\|\widehat{\beta} - \beta\|_1 = O_{\mathbb{P}}(1/\sqrt{n}) \quad \textit{and} \quad \|\widehat{\gamma} - \gamma\|_1 = O_{\mathbb{P}}(1/\sqrt{n}), \tag{23}$$

then Theorem 2's conclusion holds and $\phi_{n,\alpha}^{\text{spaCRT}}$ *controls Type-I error asymptotically.*

Note that the $\sqrt{n}$ rate condition in (23) is classical for MLEs under mild conditions.

High-dimensional regression. We now demonstrate how spaCRT can be used in the presence of high-dimensional parameters, so we allow dimension of $\boldsymbol{Z}$ (as well as β_n) to grow with sample size n. In particular, we consider the estimators $\widehat{\beta}_n, \widehat{\gamma}_n$ for β_n, γ_n obtained from the lasso estimators (Tibshirani, 1996) with regularization parameters λ_n, ν_n, respectively. The definitions of these estimators are standard and can be found in Appendix M. We show that the spaCRT is asymptotically valid if β_n and γ_n are sparse enough, and if the following assumption on the covariate distribution $\boldsymbol{Z}$ holds.

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

$$\inf_n \lambda_{\min}(\mathbb{E}_{\mathcal{L}_n}[\boldsymbol{Z}\boldsymbol{Z}^\top]) > 0; \tag{24}$$

$$\sup_n \sup_{\boldsymbol{\eta}\in\mathbb{R}^d, \|\boldsymbol{\eta}\|_2=1} \mathbb{E}_{\mathcal{L}_n}[\langle \boldsymbol{\eta}, \boldsymbol{Z}\rangle^4] < \infty. \tag{25}$$

Theorem 4. *Suppose $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ and Assumptions 3-5 hold. Defining $(s_{\beta_n}, s_{\gamma_n}) \equiv (\|\beta_n\|_0, \|\gamma_n\|_0)$, suppose there exists $\delta \in (0,1)$ such that*

$$\max\{1, s_{\gamma_n}, s_{\beta_n}\}\sqrt{\log(d)/n} \asymp n^{-\delta}; \tag{26}$$

$$\sup_n \|\gamma_n\|_1 < \infty, \ \sup_n \|\beta_n\|_1 < \infty. \tag{27}$$

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 of Theorem 2 holds. If, additionally, we have

$$s_{\gamma_n} \cdot s_{\beta_n} \cdot \frac{\log(d)}{n^{1/2}} = o(1), \tag{28}$$

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

Let us comment on the assumptions required. Assumption 5 imposes constraints on the covariate distribution, which are commonly required in high-dimensional regression (Wainwright, 2019). Condition (26) regulates the growth rates of $s_{\gamma_n}, s_{\beta_n}, d$ and n. Note that the $\asymp$ symbol in this condition denotes the same order of growth; it is formally defined in Appendix A.1. The boundedness condition (27) is a relatively mild condition required to show the almost sure convergence of $\widehat{\mu}_{n,y}(\cdot), \widehat{\theta}_{n,x}(\cdot)$ in verifying condition (19). Condition (28) requires that the product of the sparsity levels of β_n and γ_n is small enough, reflecting the double robustness of spaCRT (Remark 7).

5 Application to simulated data with high sparsity

As discussed in Section 1, the spaCRT is particularly useful for the analysis of sparse data. We demonstrate the advantages of spaCRT in this context via two simulation studies, inspired by single-cell CRISPR screens analysis (Section 5.1) and GWAS for a rare disease and rare genetic variants (Section 5.2). We present additional simulation results on unbalanced classification using random forests in Appendix E.3.1. Code to reproduce all analyses in this section and the next is available at github.com/Katsevich-Lab/spacrt-manuscript.

5.1 Single-cell CRISPR screens analysis

Single-cell CRISPR screens pair genetic perturbations with single-cell gene expression measurements to identify the effects of perturbed genomic elements on gene expression (Dixit et al., 2016; Adamson et al., 2016). The perturbation presence and gene expression data produced by these assays tend to be sparse (Appendix P.1), creating the analysis challenges described in the introduction.

Simulation setup: While single-cell CRISPR screens can simultaneously measure the effects of thousands of perturbations on tens of thousands of genes, we focus our simulation study on testing for association within a single perturbation-gene pair. In conjunction with using the stricter significance threshold of $\alpha = 0.005$ that could arise from a multiplicity correction, this setup already captures the relevant statistical phenomena of interest (see also Appendix P.3 for results in a multiple testing setting). To simulate single-cell CRISPR screen data for a given perturbation-gene pair, let $\boldsymbol{X} \in \{0, 1\}$, $\boldsymbol{Y} \in \mathbb{N}$, and $\boldsymbol{Z} \in \mathbb{R}$ represent the indicator of perturbation presence, gene expression, and a single covariate with a confounding effect in a given cell. We observe these variables in each of n cells. We model $\boldsymbol{X} \mid \boldsymbol{Z}$ and $\boldsymbol{Y} \mid \boldsymbol{Z}$ as logistic and negative binomial regressions, respectively (Gasperini et al., 2019; Barry et al., 2021; Barry et al., 2024). 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; Svensson, 2020; Sarkar and Stephens, 2021). We arrive at the model

$$\boldsymbol{Z} \sim N(0, 1); \;\; \boldsymbol{X} \mid \boldsymbol{Z} \sim \text{Ber}(\text{logit}(\mu) = \gamma_0 + \boldsymbol{Z}); \tag{29}$$

$$\boldsymbol{Y} \mid \boldsymbol{X}, \boldsymbol{Z} \sim \text{NB}(\log(\mu) = \beta_0 + \rho \boldsymbol{X} + \boldsymbol{Z}, r), \tag{30}$$

where $r > 0$ is the *size parameter* controlling the overdispersion of the negative binomial distribution. 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 parameter ρ controls the strength of the signal, i.e., the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$ conditionally on $\boldsymbol{Z}$. Therefore, $\rho = 0$ and $\rho \neq 0$ correspond to the null and alternative hypotheses, respectively. We note that selecting the parameters (γ_0, β_0) yields X and Y whose sparsity levels closely resemble those observed in real data (see Appendix P.2 for details).

Methodologies compared: We compare four tests: spaCRT, dCRT, GCM test and the negative binomial score test (we defer detailed method definitions to Appendix P.2). We use $M = 10,000$ resamples for dCRT to obtain accurate p-values.

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 $\alpha = 0.005$. Additional results are provided in Appendix P.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. From Figure 2d, we see the p-values from spaCRT and dCRT align very well for p-values estimated accurately enough by $M = 10,000$ dCRT resamples, validating

our approximation accuracy result (Theorem 2) in finite samples. For very small p-values (those around 10^{-4} or smaller), the dCRT does not yield reliable estimates due to discreteness, whereas spaCRT captures p-values as small as 10^{-7}. 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.

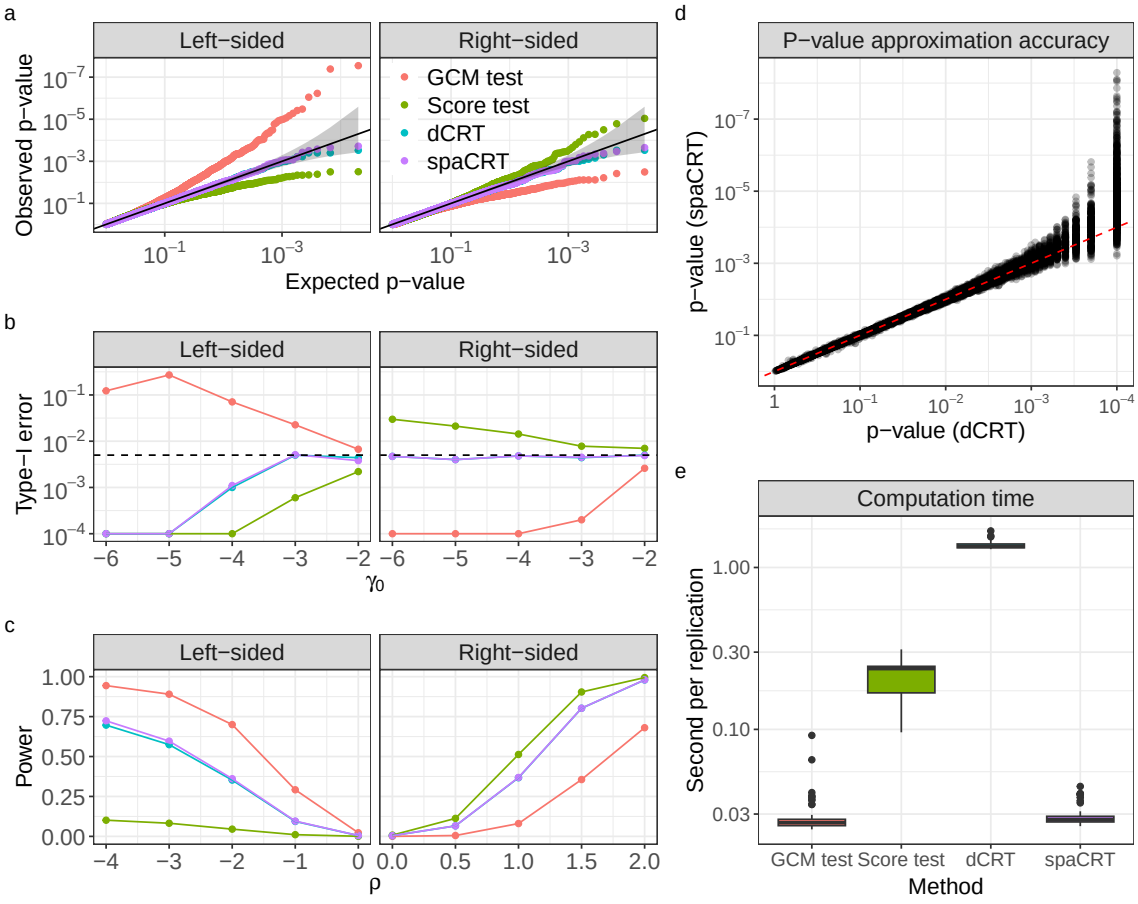

Figure 2: Summary of single-cell CRISPR screen simulation results for size parameter $r = 0.05$ and the significance level $\alpha = 0.01$. (a) QQ-plots of the p-values obtained under the null hypothesis for $(\gamma_0, \beta_0) = (-3, -5)$. (b) Type-I error rates for $\beta_0 = -5$ as a function of the sparsity of X (γ_0). (c) Power for $(\gamma_0, \beta_0) = (-3, -5)$ as a function of the signal strength (ρ). (d) Scatter plot for p-values comparison between dCRT and spaCRT for $(\gamma_0, \beta_0, \rho) = (-3, -5, 1)$. (e) Time consumed by each method across different simulation parameters. Each point in panels (b), (c), and (e) is an average over 10,000 replicates.

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 discussion on size parameter r to Appendix P.3, where we show that dCRT and spaCRT are less sensitive to this parameter than the GCM and score tests.

Finally, Figure 2e 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 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.

5.2 GWAS with rare diseases and genetic variants

The goal of genome-wide association studies (GWAS) is to identify genetic variants associated with phenotypes, such as diseases. When the genetic variants and/or the disease under investigation are rare, this makes the data highly sparse (Appendix Q.1).

Simulation setup: To simulate GWAS data, consider a binary disease indicator $\boldsymbol{Y}$ and a high-dimensional vector of genotypes $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_d)^\top \in \{0,1\}^d$ at d genetic locations. While genotypes typically take values in $\{0,1,2\}$, we simplify the setup while preserving its qualitative nature by considering binary genotypes. For each location j, it is of interest to test for association between that location's genotype and the disease, controlling for the effects of other genetic variants. This variable selection problem has been formulated as a collection of CI testing problems (Sesia, Sabatti, and Candès, 2019) fitting into our framework by considering $\boldsymbol{Z} \equiv \boldsymbol{X}_{-j}$ as the covariates:

$$H_0^j : \boldsymbol{X}_j \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{X}_{-j}, \quad j = 1, \ldots, d. \tag{31}$$

We model the response as a high-dimensional logistic regression:

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

Under this model, the CI null H_0^j (31) is equivalent to $\beta_j = 0$ (Candès et al., 2018). γ_0 controls the sparsity of outcome Y and we consider $\{-3, -2\}$ for *high* and *low* sparsity (i.e. more rare and less rare disease). We model the genotype vector $\boldsymbol{X} \in \{0,1\}^d$ as a *hidden Markov model* (HMM), commonly adopted for this purpose (Scheet, Stephens, and Scheet, 2006; Marchini et al., 2007; Browning and Browning, 2007) and reviewed in Appendix B. We consider two sets of HMM parameters, capturing rare and common genetic variation. We set dimension $d = 500$ and sample size $n = 2000$, and describe the other parameter settings to Appendix Q.2.

Methodologies compared: We compare four procedures: spaCRT, dCRT, GCM test and the knockoff procedure (Candès et al., 2018; Sesia, Sabatti, and Candès, 2019). All methods use the expectation-maximization (EM) algorithm of Scheet, Stephens, and Scheet (2006) to fit the HMM distribution of $\boldsymbol{X}$ and lasso-regularized logistic regression to fit $\boldsymbol{Y}|\boldsymbol{X}$. The method details can be found in Appendix Q.2. We apply a two-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 false discovery rate (FDR) is controlled at level $q = 0.1$.

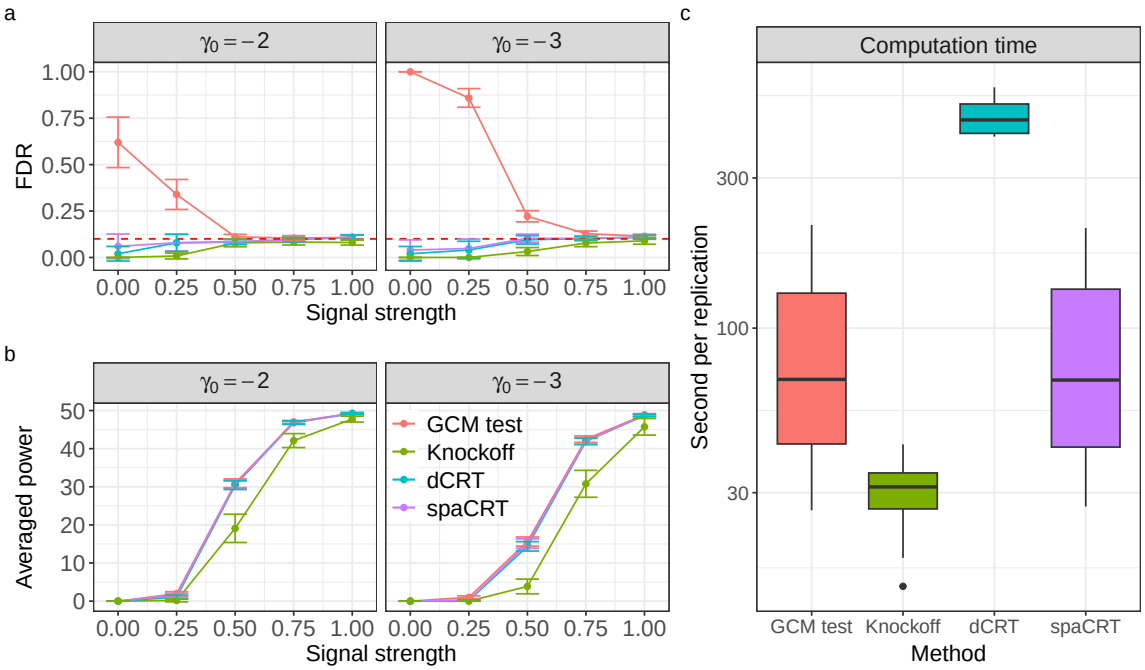

Figure 3: Summary of GWAS simulation results with high sparsity in data X. All the results are obtained with the regularization parameter $\lambda =$ `lambda.1se`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Time consumed by different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

Simulation results: We present a representative selection of simulation results (Figure 3) with high sparsity in X; the other results can be found in Appendix Q.4. Note that the knockoff procedure does not provide p-values but just a rejection set. 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 high sparsity in X and Y, although the degree of inflation is more severe when $\gamma_0 = -3$. The knockoff method also controls FDR, though is conservative when the signal is weak. Figure 3b shows that GCM, dCRT and spaCRT have similar power, whereas knockoff procedure tends to be less powerful. We conjecture the latter may be due to the sensitivity of knockoffs to the distribution of $\boldsymbol{X}$, which can affect the correlation of knockoff variables $\tilde{\boldsymbol{X}}$ with the originals $\boldsymbol{X}$, which in turn affects power. Therefore, we do not claim that the knockoff procedure is generally less powerful and postpone a more comprehensive investigation to future work.

Figure 3c shows the time consumption for different methods. Among all methods, knockoffs is the most computationally efficient, due to the fact that only one high-dimensional regression fit of $\boldsymbol{Y}|\boldsymbol{X}, \tilde{\boldsymbol{X}}$ is required. 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 resampling required by dCRT. However, we employ a *tower trick* (Chakraborty, Zhang, and Katsevich, 2024) to accelerate the computation of $\widehat{\mathbb{E}}[\boldsymbol{Y}|\boldsymbol{X}_{-j}]$ for dCRT, GCM and spaCRT (Appendix Q.3). This brings the computation time of spaCRT and GCM within a factor of 2 to 3 of the knockoff procedure while being at

least 4 times faster than the dCRT. Part of the remaining advantage of knockoffs is due to its highly optimized implementation in the R package `SNPknock`. Finally, we note that the computational advantage of spaCRT over dCRT is not as dramatic due to the relatively small multiplicity in our simulation ($d = 500$), which allows us to set $M = 5000$. Full-scale GWAS analyses involve millions of genetic variants, requiring a much larger M and significantly increasing the per-test computational cost of dCRT.

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 R.1 for a 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 use negative and positive control CRISPR perturbations, 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. To assess power, we test for association between each of the 754 positive control perturbations and the genes they target.

Methods compared. We compare essentially the same methods as in the numerical simulations (recall Section 5.1), except we replace the dCRT with a faster variant implemented in the R package `sceptre` (Barry et al., 2021; Barry et al., 2024), in order to make the analysis computationally feasible. As discussed in Section 1.2, 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. 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` p-values are well-calibrated for both left- and right-sided tests. We report the number of false discoveries on the negative control pairs in Appendix R.3; these results are what one would expect based on the QQ plots.

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

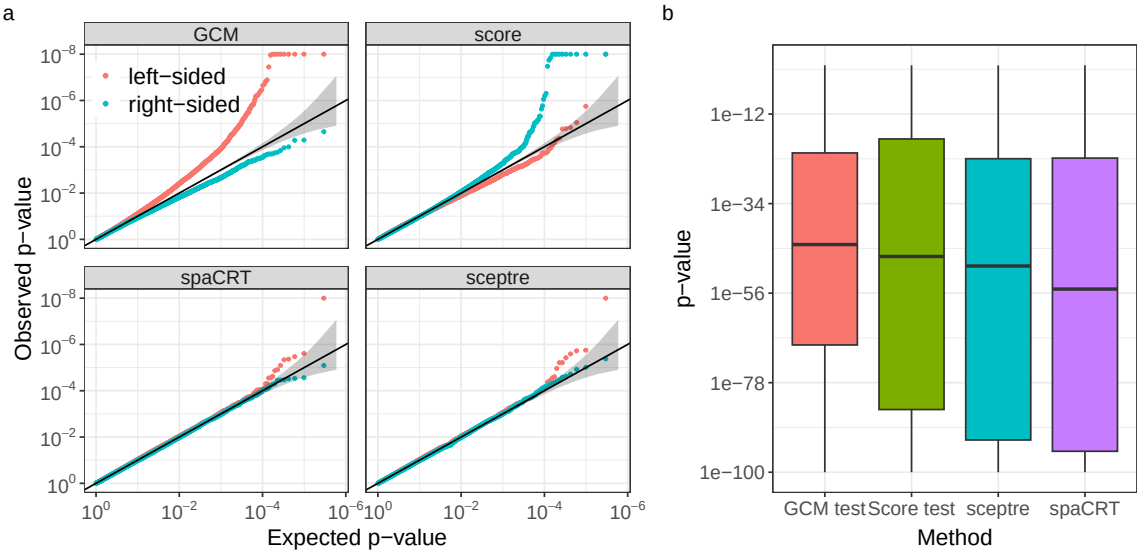

Figure 4: Calibration and power performance on the Gasperini et al. (2019) data. (a): Left- and right-sided p-values for negative control perturbation-gene pairs. (b): Left-sided p-values computed on the positive control perturbation-gene pairs.

nonzero expression of a given gene. Table 5 in Appendix R.2 displays the distribution of effective sample sizes across the negative control pairs tested, showing that the effective samples sizes are vastly smaller than the number of cells, $n = 207{,}324$. Furthermore, Figure 17 in Appendix R.4 stratifies the QQ plots for each method by effective sample size, 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 also stratified pairs based on the estimated size parameter (Figure 18 in Appendix R.4). As in our simulation results, the GCM and score tests exhibit more miscalibration for smaller size parameters.

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. In addition to its excellent statistical performance, spaCRT enjoys excellent computational performance as well. Since the `sceptre` software is highly optimized, while the other methods are not, we benchmark the computational cost of our dCRT implementation (used in the numerical simulations) instead of `sceptre`'s. We use $M = 100{,}000$ resamples for the dCRT, given the high multiplicity of the problem. We assess runtime on 102 pairs based on two randomly sampled genes and

the 51 non-targeting perturbations. We find that the spaCRT is roughly as fast as the GCM test, five times faster than the score test, and 250 times faster than the dCRT.

Table 1: Computation times (in seconds) per test on the Gasperini data.

Method	Mean	Std dev
GCM test	2.4	1.2
dCRT	566.4	48.7
spaCRT	2.4	1.0
Score test	12.8	2.6

7 Discussion

In this paper, we demonstrated that the saddlepoint approximation (SPA), previously regarded as a classical and somewhat specialized tool, can be effectively adapted to address modern, large-scale statistical inference challenges. Our primary contributions are twofold. First, we established a general conditional Lugannani–Rice result with vanishing relative error under transparent, easily verifiable tail conditions. Our approach fills decades-old theoretical gaps in the justification of SPAs for resampling-based procedures and unifies previously separate smooth and lattice cases. Second, leveraging this theoretical foundation, we developed the spaCRT, a resampling-free CI test. The spaCRT retains the excellent finite-sample statistical performance of the dCRT while reducing computational requirements by orders of magnitude (on the Gasperini data, the reduction was from years to days). Practically, researchers conducting large-scale hypothesis testing on sparse data now have access to a method (immediately deployable via the open-source `spacrt` package) that balances computational efficiency with statistical accuracy, providing reliably calibrated p-values without the computational overhead of resampling-based procedures.

Our work opens a number of directions for future research. Theorem 1 can already be applied to establish conditions for the validity of existing SPAs to resampling-based hypothesis tests with conditionally independent summands, such as bootstrap tests. This result, though already quite general, can be extended in at least three ways. First, it can be extended to exchangeable yet conditionally dependent resampling schemes, in order to cover permutation tests. Second, it can be generalized to accommodate fixed cutoff sequences (in addition to its current scope of vanishing cutoff sequences), which would strengthen its applicability in extreme-tail inference. Third, it can be generalized to accommodate standardized test statistics, which are known to improve the performance of resampling-based methods (Hall, 2010; Chung and Romano, 2013; Barry et al., 2025). With some of these extensions, the SPA can be applied to a wider range of resampling-based procedures, including the CRT with a broader class of test statistics. Finally, an adjacent question that remains open is to theoretically justify the empirically observed finite-sample improvement of the dCRT over asymptotic variants like the GCM test (Theorem 7 in Appendix D.2 proves this only in a special case).

In conclusion, by establishing the rigorous theoretical underpinnings for the conditional SPA and embedding it within a versatile conditional independence testing frame-

work, our work significantly advances the methodology available for high-dimensional, sparse inference problems. The spaCRT represents a substantial step forward, combining computational efficiency with rigorous statistical guarantees, thus opening avenues for further methodological development and broader application.

8 Acknowledgments

We are very grateful to John Kolassa, who provided valuable feedback on an earlier version of this paper. 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 supported 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.

Barber, Rina Foygel and Emmanuel J. Candés (2015). "Controlling the false discovery rate via knockoffs". In: *Annals of Statistics* 43.5, pp. 2055–2085. arXiv: 1404.5609.

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, Ziang Niu, Eugene Katsevich, and Xihong Lin (2025). "The permuted score test for robust differential expression analysis". In: *arXiv*.

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". In: *Source: Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300.

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

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: *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.

Chakraborty, Abhinav, Jeffrey Zhang, and Eugene Katsevich (2024). "Doubly robust and computationally efficient high-dimensional variable selection". In: arXiv: 2409.09512v1.

Chen, Louis H.Y., Larry Goldstein, and Qi-Man Shao (2011). *Normal approximations by Stein's method*. Vol. 23. 1. Springer.

Chung, Eunyi and Joseph P. Romano (2013). "Exact and asymptotically robust permutation tests". In: *Annals of Statistics* 41.2, pp. 484–507.

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.

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: *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 (1979). "Bootstrap Methods: Another Look at the Jackknife". In: *Annals of Statistics* 7.1, pp. 1–26.

— (2022). *Exponential Families in Theory and Practice*. Cambridge University Press, pp. 1–250.

Fischer, Lasse, Timothy Barry, and Aaditya Ramdas (2024). "Multiple testing with anytime-valid Monte-Carlo p-values". In: pp. 1–22. arXiv: 2404.15586.

Fischer, Lasse and Aaditya Ramdas (2024). "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: *Source: 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 (2010). *The Bootstrap and the Edgeworth expansion*.

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.

Klenke, Achim (2017). *Probability theory*. Vol. 941, pp. 1–23.

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.

Koller, Daphne and Nir Friedman (2009). *Probabilistic Graphical Models*. MIT Press.

Kuchibhotla, Arun Kumar (2023). "Central Limit Theorems and Approximation Theory: Part II". In: arXiv: 2306.14382.

Lauritzen, Steffen L. (1996). *Graphical models*. Clarendon Press.

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.

Magwene, Paul M. and Junhyong Kim (2004). "Estimating genomic coexpression networks using first-order conditional independence." In: *Genome biology* 5.12.

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*.

Pearl, Judea (2009). "Causal inference in statistics: An overview". In: *Statistics Surveys* 3, pp. 96–146.

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

Reid, N. (1988). "Saddlepoint methods and statistical inference". In: *Statistical Science* 3.2, pp. 213–238. arXiv: arXiv:1011.1669v3.

Robins, James M. and Andrea Rotnitzky (2001). "Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer"". In: *Statistica Sinica* 11.4, pp. 920–936.

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.

Sarkar, Abhishek and Matthew Stephens (2021). "Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis". In: *Nature Genetics*.

Scheet, Paul, Matthew Stephens, and Mr Paul Scheet (2006). "A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to In-

ferring Missing Genotypes and Haplotypic Phase". In: *The American Journal of Human Genetics* 78, pp. 629–644.
Sekulovski, Nikola et al. (2024). "Testing Conditional Independence in Psychometric Networks: An Analysis of Three Bayesian Methods". In: *Multivariate Behavioral Research* 59.5, pp. 913–933.
Sesia, M., C. Sabatti, and E. 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.
Smucler, Ezequiel, Andrea Rotnitzky, and James M. Robins (2019). "A unifying approach for doubly-robust L1 regularized estimation of causal contrasts". In: *arXiv*. arXiv: 1904.03737.
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: *J. R. Statist. Soc. B* 58.1, pp. 267–288.
Wainwright, Martin J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge university press.
Wang, Zhu, Shuangge Ma, and Ching Yun Wang (2015). "Variable selection for zero-inflated and overdispersed data with application to health care demand in Germany". In: *Biometrical Journal* 57.5, pp. 867–884.
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.
Yao, Sirui and Bert Huang (2017). "Beyond parity: Fairness objectives for collaborative filtering". In: *Advances in Neural Information Processing Systems*. Vol. 2017-Decem, pp. 2922–2931. arXiv: 1705.08804.
Zhao, Zhangchen et al. (2020). "UK Biobank Whole-Exome Sequence Binary Phenome Analysis with Robust Region-Based Rare-Variant Test". In: *American Journal of Human Genetics* 106.1, pp. 3–12.

Appendix

Notations. Suppose f is an infinitely differentiable function and we will use $f^{(r)}$ to denote its r-th derivative.

Index in the supplement. We will use prefix $S.$ to denote the section and theoretical results number and "E.a" to denote the equation number. For example "Section S.1" or "Conclusion (E.1)". When there is no prefix in the index, it means the section or equation is in the main paper. For example, "Section 2" or "Equation (1)".

A Probability theory preliminaries

A.1 Asymptotic notations

We use the following standard notations regarding the asymptotic properties of a sequence of random variables 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 will write $a_n \asymp b_n$ if $0 < \liminf_{n\to\infty} |a_n/b_n| \le \limsup_{n\to\infty} |a_n/b_n| < \infty$ and $a_n = o(b_n)$ if $\limsup_{n\to\infty} |a_n/b_n| = 0$. Moreover, we will write $a_n \lesssim b_n$ if there exists constant C such that $|a_n/b_n| \le C$ when n is large enough.

A.2 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 $\{(\mathbb{P}_n, \Omega_n, \mathcal{G}_n), n \ge 1\}$. For each n, let $\{W_{i,n}\}_{i\ge1}$ 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 $\{\widetilde{W}_{i,n}\}_{i,n\ge1}$ on $(\widetilde{\mathbb{P}}, \widetilde{\Omega}, \widetilde{\mathcal{G}})$, and σ-fields $\widetilde{\mathcal{F}}_n \subseteq \widetilde{\mathcal{G}}$ for $n \ge 1$, such that for each n, the joint distribution of $(\{W_{i,n}\}_{i\ge1}, \{\mathbb{E}[W_{i,n} \mid \mathcal{F}_n]\}_{i\ge1})$ on $(\mathbb{P}_n, \Omega_n, \mathcal{G}_n)$ coincides with that of $(\{\widetilde{W}_{i,n}\}_{i\ge1}, \{\mathbb{E}[\widetilde{W}_{i,n} \mid \widetilde{\mathcal{F}}_n]\}_{i\ge1})$ 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.3 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. $\mathbb{E}[(X - \mathbb{E}[X])^4] - 3\left(\mathbb{E}[(X - \mathbb{E}[X])^2]\right)^2 = A^{(4)}(\theta)$.

A.4 Preliminaries on regular conditional distribution

To better understand the argument involving conditional distribution, we briefly discuss the basic definition of regular conditional distribution (RCD). Let $\mathcal{B}(\mathbb{R}^n)$ be the Borel σ-algebra on $\mathbb{R}^n$ and $\Omega, \mathcal{F}_n$ be the sample space and a sequence of σ-algebras. For any $n \in \mathbb{N}_+, \kappa_n : \Omega \times \mathcal{B}(\mathbb{R}^n)$ is a regular conditional distribution of $W_n \equiv (W_{1n}, \ldots, W_{nn})^\top$ given $\mathcal{F}_n$ if

$$\begin{aligned} \omega \mapsto \kappa_n(\omega, B) \text{ is measurable with respect to } \mathcal{F}_n \text{ for any fixed } B \in \mathcal{B}(\mathbb{R}^n); \\ B \mapsto \kappa_n(\omega, B) \text{ is a probability measure on } (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n)); \\ \kappa_n(\omega, B) = \mathbb{P}[(W_{1n}, \ldots, W_{nn}) \in B | \mathcal{F}_n](\omega), \text{ for almost all } \omega \in \Omega \text{ and all } B \in \mathcal{B}(\mathbb{R}^n). \end{aligned}$$

The following lemma from Klenke, 2017, Theorem 8.37 ensures that the general existence of κ_n.

Lemma 3 (Theorem 8.37 in Klenke, 2017)**.** *Suppose $(\Omega, \mathcal{G}, \mathbb{P})$ is the Probability triple. Let $\mathcal{F} \subset \mathcal{G}$ be a sub-σ-algebra. Let Y be a random variable with values in a Borel space $(E, \mathcal{E})$ (for example, E is Polish, $E = \mathbb{R}^d$). Then there exists a regular conditional distribution $\kappa_{Y,\mathcal{F}}$ of Y given $\mathcal{F}$.*

Result from Klenke, 2017, Theorem 8.38 guarantees that the conditional expectation and the integral of measurable function with respect to regular conditional distribution are almost surely same.

Lemma 4 (Modified version of Theorem 8.38 in Klenke, 2017)**.** *Let X be a random variable $(\Omega, \mathcal{G}, \mathbb{P})$ with values in a Borel space $(E, \mathcal{E})$. Let $\mathcal{F} \subset \mathcal{G}$ be a σ-algebra and let $\kappa_{X,\mathcal{F}}$ be a version of regular conditional distribution of X given $\mathcal{F}$. Further, let $f : E \to \mathbb{R}$ be measurable and $\mathbb{E}[|f(X)|] < \infty$. Then we can define a version of the conditional expectation of $f(X)$ given $\mathcal{F}$ as:*

$$\mathbb{E}[f(X)|\mathcal{F}](\omega) = \int f(x) \mathrm{d}\kappa_{X,\mathcal{F}}(\omega, x), \ \forall \omega \in \Omega.$$

A.5 A version of conditional expectation

In this paper, we will use RCD to fix a version of the conditional expectations used in the paper. Suppose $\kappa_{in} : \Omega \times \mathcal{B}(\mathbb{R})$ is a version of the RCD of W_{in} given $\mathcal{F}_n$ and $\kappa_n : \Omega \times \mathcal{B}(\mathbb{R}^n)$ is a version of the RCD of $W_n \equiv (W_{1n}, \ldots, W_{nn})^\top$ given $\mathcal{F}_n$. The conditional independent assumption can be formulated as

$$\kappa_n(\omega, B) = \prod_{i=1}^{n} \kappa_{in}(\omega, B_i), \ B = B_1 \times \ldots \times B_n, \ \forall B_1, \ldots, B_n \in \mathcal{B}(\mathbb{R}), \ \forall \omega \in \Omega.$$

Define the tilted RCD:

$$\frac{\mathrm{d}\kappa_{in,s}(\omega, x)}{\mathrm{d}\kappa_{in}(\omega, x)} \equiv \frac{\exp(sx)}{\int \exp(sx)\mathrm{d}\kappa_{in}(\omega, x)}, \ \frac{\mathrm{d}\kappa_{n,s}(\omega, x)}{\mathrm{d}\kappa_{n}(\omega, x)} \equiv \prod_{i=1}^{n} \frac{\exp(sx)}{\int \exp(sx)\mathrm{d}\kappa_{in}(\omega, x)}, \ \forall \omega \in \Omega.$$

Given tilting parameter s, define measure $\mathbb{P}_{in,s}$ and $\mathbb{P}_{n,s}$ on the measurable space $(\Omega, \mathcal{F})$ via

$$\frac{\mathrm{d}\mathbb{P}_{in,s}}{\mathrm{d}\mathbb{P}} \equiv \frac{\exp(sW_{in})}{\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n]} \quad \text{and} \quad \frac{\mathrm{d}\mathbb{P}_{n,s}}{\mathrm{d}\mathbb{P}} \equiv \prod_{i=1}^{n} \frac{\mathrm{d}\mathbb{P}_{in,s}}{\mathrm{d}\mathbb{P}}.$$

For any measurable function $f : \mathbb{R} \mapsto \mathbb{R}$ and $g : \mathbb{R}^n \mapsto \mathbb{R}$, we define the conditional expectation under original measure $\mathbb{P}$, tilted measure $\mathbb{P}_{in,s}$ and $\mathbb{P}_{n,s}$ respectively as

$$\mathbb{E}[f(W_{in})|\mathcal{F}_n](\omega) \equiv \int f(x)\mathrm{d}\kappa_{in}(\omega, x), \ \forall \omega \in \Omega, \tag{32}$$

$$\mathbb{E}_{in,s}[f(W_{in})|\mathcal{F}_n](\omega) \equiv \int f(x)\frac{\exp(sx)}{\int \exp(sx)\mathrm{d}\kappa_{in}(\omega, x)}\mathrm{d}\kappa_{in}(\omega, x), \ \forall \omega \in \Omega, \tag{33}$$

$$\mathbb{E}_{n,s}[g(W_{n})|\mathcal{F}_n](\omega) \equiv \int g(y)\left(\prod_{i=1}^{n}\frac{\exp(sy_i)}{\int \exp(sy_i)\mathrm{d}\kappa_{in}(\omega, y_i)}\right)\mathrm{d}\kappa_{n}(\omega, y), \ \forall \omega \in \Omega \tag{34}$$

where $x \in \mathbb{R}, y \in \mathbb{R}^n$. The above results provide a version of conditional expectation via RCD. We refer the guarantee of existence of RCD and the validity of the above definition for conditional expectation to Theorem 3 and 4.

B An introduction to HMM

B.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 $\boldsymbol{X} \in \{0, 1\}^p$. 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* $\boldsymbol{U} \in \{0, 1, \ldots, K\}^p$ *follows a multinomial Markov Chain distribution,* $\boldsymbol{U} \equiv (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_p)^\top \sim mMC(q, Q)$*, 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* $\{0, 1, \ldots, K\}$ *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* $u \in \{1, \ldots, K\}$*.*

Then we consider the following definition for mHMM:

Definition 2 (mHMM)**.** *Consider a p-dimensional random variable* $\boldsymbol{X} \in \{0, 1\}^p$*. 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.

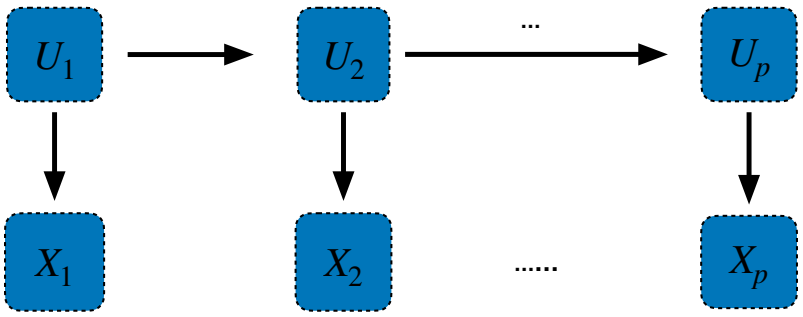

Figure 5: Graphical illustration of a multinomial hidden Markov model. Dashed rounded rectangles: unobserved variables.

The following identities can be easily obtained:

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

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

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 $\boldsymbol{X}_{1:j}$.

B.2 Computing conditional distribution-related quantities

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 computed. We first boil these quantities down to the conditional probability $\mathbb{P}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = x_{-j}]$, where $x_{-j} \equiv (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)^\top \in \mathbb{R}^{p-1}$. Then we will show how the conditional probability can be computed using forward-backward algorithm. For the ease of notation, we now define for $x_{-j} \in \mathbb{R}^{p-1}$ and $t \in \mathbb{R}$,

$$p(x_j, x_{-j}) \equiv \mathbb{P}[\boldsymbol{X}_j = x_j|\boldsymbol{X}_{-j} = x_{-j}] \quad \text{and} \quad D(t, x_{-j}) \equiv \mathbb{E}[\exp(t\boldsymbol{X}_j)|\boldsymbol{X}_{-j} = x_{-j}].$$

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. Then 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:

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

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 $x \in \{0, 1\}^p$, we define

$$A_j(u, x) \equiv \mathbb{P}[\boldsymbol{X}_{1:(j-1)} = x_{1:(j-1)}, \boldsymbol{U}_j = u], \ A_1(u, x) \equiv \mathbb{P}[\boldsymbol{U}_1 = u]; \tag{35}$$

$$B_j(u, x) \equiv \mathbb{P}[\boldsymbol{X}_{(j+1):p} = x_{(j+1):p}|\boldsymbol{U}_j = u], \ B_p(u, x) \equiv 1. \tag{36}$$

Proposition 1 (Iterative computation of $p(x_j, x_{-j})$)**.** *Suppose $\boldsymbol{X} \sim mHMM(\boldsymbol{U}(q, Q), e)$. Then we have*

$$p(x_j, x_{-j}) = \frac{\sum_{\bar{u} \in \{0,1,\ldots,K\}} e(x_j|\bar{u}) \cdot A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}{\sum_{\bar{u} \in \{0,1,\ldots,K\}} A_j(\bar{u}, x) \cdot B_j(\bar{u}, x)}. \tag{37}$$

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

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

The proof can be found in Appendix B.3.

B.3 Proof of Proposition 1

Proof of conclusion (37). First, let us consider the following marginalization:

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

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

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

Thus we have

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

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

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

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:

$$\begin{aligned}
A_2(u,x) &= \mathbb{P}[\boldsymbol{X}_1 = x_1, \boldsymbol{U}_2 = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_1 = \bar{u}]\mathbb{P}[\boldsymbol{X}_1 = x_1, \boldsymbol{U}_2 = z|\boldsymbol{U}_1 = \bar{u}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{U}_1 = \bar{u}]\mathbb{P}[\boldsymbol{X}_1 = x_1|\boldsymbol{U}_1 = \bar{u}]\mathbb{P}[\boldsymbol{U}_2 = u|\boldsymbol{U}_1 = \bar{u}] \quad \text{(CI property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} A_1(\bar{u},x)\cdot e(x_1|\bar{u})\cdot Q(u|\bar{u}). \qquad \text{(Definition (35))}
\end{aligned}$$

Then we can easily show that

$$\begin{aligned}
A_j(u,x) &= \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&\qquad \times \mathbb{P}[\boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{Z}_j = z|\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_1 = x_1, \ldots, \boldsymbol{X}_{j-2} = x_{j-2}, \boldsymbol{U}_{j-1} = \bar{u}] \\
&\qquad \times \mathbb{P}[\boldsymbol{X}_{j-1} = x_{j-1}, \boldsymbol{U}_j = u|\boldsymbol{U}_{j-1} = \bar{u}] \qquad \text{(Markov property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} A_{j-1}(\bar{u},x)\cdot e(x_{j-1}|\bar{u})\cdot Q(u|\bar{u}). \qquad \text{(Definition (35))}
\end{aligned}$$

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

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

$$\begin{aligned}
B_{p-1}(u,x) &= \mathbb{P}[\boldsymbol{X}_p = x_p|\boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p, \boldsymbol{U}_p = \bar{u}|\boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p|\boldsymbol{U}_p = \bar{u}, \boldsymbol{U}_{p-1} = u]\mathbb{P}[\boldsymbol{U}_p = \bar{u}|\boldsymbol{U}_{p-1} = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_p = x_p|\boldsymbol{U}_p = \bar{u}]\mathbb{P}[\boldsymbol{U}_p = \bar{u}|\boldsymbol{U}_{p-1} = u] \quad \text{(Markov property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} B_p(\bar{u},x)\cdot e(x_p|\bar{u})\cdot Q(\bar{u}|u). \qquad \text{(Definition (36))}
\end{aligned}$$

By induction, we can compute

$$\begin{aligned}
B_j(u,x) &= \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \ldots, \boldsymbol{X}_p = x_p, \boldsymbol{U}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+2} = x_{j+2}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_{j+1} = \bar{u}, \boldsymbol{U}_j = u, \boldsymbol{X}_{j+1} = x_{j+1}] \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1}, \boldsymbol{Z}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} \mathbb{P}[\boldsymbol{X}_{j+2} = x_{j+2}, \ldots, \boldsymbol{X}_p = x_p | \boldsymbol{U}_{j+1} = \bar{u}] \qquad \text{(Markov property)} \\
&\quad \times \mathbb{P}[\boldsymbol{X}_{j+1} = x_{j+1} | \boldsymbol{U}_{j+1} = \bar{u}] \cdot \mathbb{P}[\boldsymbol{U}_{j+1} = \bar{u} | \boldsymbol{U}_j = u] \qquad \text{(CI property)} \\
&= \sum_{\bar{u}\in\{0,1,\ldots,K\}} B_{j+1}(\bar{u}, x) \cdot e(x_{j+1}|\bar{u}) \cdot Q(\bar{u}|u). \qquad \text{(Definition (36))}
\end{aligned}$$

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

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 particular, fixing a set of value x_{-j}, we compute

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

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

C Additional details of Section 2

C.1 Additional details of Section 2.1

We first discuss the existence of the conditional cumulant generating function $K_{in}(s)$ and its derivatives. Either Assumption 1 or Assumption 2 guarantees the existence of the CGFs $K_{in}(s)$ and their derivatives in a neighborhood of the origin. This is formalized in the following lemma.

Lemma 5. *Suppose Assumption 1 or Assumption 2 holds. Then, there exists a probability-one event $\mathcal{A}$ and an $\varepsilon > 0$ such that, on $\mathcal{A}$,*

$$K_{in}(s) < \infty \quad \textit{for any } s \in (-\varepsilon, \varepsilon) \textit{ and for all } i \le n,\ n \ge 1 \tag{38}$$

and

$$|K_{in}^{(r)}(s)| < \infty \quad \textit{for any } s \in (-\varepsilon, \varepsilon) \textit{ and for all } i \le n,\ n \ge 1,\ r \ge 1, r \in \mathbb{N}, \tag{39}$$

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

We now explicitly define the solution to saddlepoint equation (4) $\hat{s}_n$. In particular, consider ε given in Lemma 5, 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:

$$\hat{s}_n \equiv \begin{cases} \text{the single element of } S_n & \text{if } |S_n| = 1; \\ \frac{\varepsilon}{2}\text{sgn}(w_n) & \text{otherwise.} \end{cases} \tag{40}$$

Note that this definition ensures that $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$. Also, we state a complete version of Theorem 1 below, which includes the conditions for the existence of the saddlepoint approximation. We will prove in Section G that, the saddlepoint equation (4) has a unique and finite solution $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$ with probability approaching 1 as $n \to \infty$:

$$\lim_{n\to\infty} \mathbb{P}[|S_n| = 1] = 1. \tag{41}$$

C.2 A connection to Robinson's formula (Robinson, 1982)

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 2 that employs a conditional variant of Robinson's formula:

$$\widehat{\mathbb{P}}_{\text{R}}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] \equiv \exp\left(\frac{\lambda_n^2 - r_n^2}{2}\right)(1 - \Phi(\lambda_n)). \tag{42}$$

Proposition 2. *Under the assumptions of Theorem 1,*

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{R}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \tag{43}$$

In other words, $\widehat{\mathbb{P}}_{\text{LR}}$ (6) is equivalent to $\widehat{\mathbb{P}}_{\text{R}}$ (42) 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). The proof of Proposition 2 is postponed to Section H.

C.3 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

$$X_{in} = \mu_n + \varepsilon_{in}, \quad \varepsilon_{in} \overset{\text{ind}}{\sim} F_{in}, \tag{44}$$

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

$$H_{0n} : \mu_n = 0 \quad \text{versus} \quad H_{1n} : \mu_n > 0 \tag{45}$$

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

$$\widetilde{X}_{in} \equiv \pi_{in} X_{in}, \quad \pi_{in} \overset{\text{i.i.d.}}{\sim} \text{Rad}(0.5), \tag{46}$$

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 (1) (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

$$\widetilde{T}_n \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{X}_{in} \equiv \frac{1}{n}\sum_{i=1}^n \pi_{in} X_{in}. \tag{47}$$

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

$$K'_{in}(s) = X_{in} - \frac{2X_{in}}{1+\exp(2sX_{in})} \quad \text{and} \quad K''_{in}(s) = \frac{4X_{in}^2 \exp(2sX_{in})}{(1+\exp(2sX_{in}))^2}. \tag{48}$$

Therefore, the saddlepoint equation (4) reduces to

$$\frac{1}{n}\sum_{i=1}^n K'_{in}(s) = \frac{1}{n}\sum_{i=1}^n X_i \quad \Longleftrightarrow \quad \sum_{i=1}^n \frac{X_{in}}{1+\exp(2sX_{in})} = 0. \tag{49}$$

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 (5):

$$\lambda_n \equiv \hat{s}_n\sqrt{nK''_n(\hat{s}_n)} = \hat{s}_n\sqrt{\sum_{i=1}^n \frac{4X_{in}^2\exp(2\hat{s}_nX_{in})}{(1+\exp(2\hat{s}_nX_{in}))^2}} \tag{50}$$

and

$$r_n \equiv \mathrm{sgn}(\hat{s}_n)\sqrt{2n(\hat{s}_n w_n - K_n(\hat{s}_n))} = \mathrm{sgn}(\hat{s}_n)\sqrt{2\sum_{i=1}^{n}(\hat{s}_n X_{in} - \log\cosh(\hat{s}_n X_{in}))}, \quad (51)$$

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

$$\widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^{n}X_{in} \;\middle|\; \mathcal{F}_n\right] \equiv 1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}. \quad (52)$$

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 (44), such that*

$$\liminf_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\varepsilon_{in}^2] > 0; \quad (53)$$

$$\textit{there exists } \delta > 0 \textit{ such that } \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|\varepsilon_{in}|^{4+\delta}] < \infty; \quad (54)$$

$$\mu_n = o(1). \quad (55)$$

Then, the saddlepoint equation (49) has a unique solution $\hat{s}_n \in [-1, 1]$ with probability approaching 1 as $n \to \infty$. Furthermore, the tail probability approximation (52) obtained from equations (50) and (51) has vanishing relative error:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^{n}X_{in} \mid \mathcal{F}_n\right] = \widehat{\mathbb{P}}_{\mathrm{LR}}\left[\frac{1}{n}\sum_{i=1}^{n}\widetilde{X}_{in} \geq \frac{1}{n}\sum_{i=1}^{n}X_{in} \;\middle|\; \mathcal{F}_n\right](1 + o_{\mathbb{P}}(1)). \quad (56)$$

We postpone the proof to Section J.

D Additional details of Section 3.1: dCRT versus GCM test

D.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 $T_n^{\mathrm{dCRT}}(X, Y, Z)$ (14) 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 $\widehat{\mu}_{n,y}(\cdot)$.

Rate and model double-robustness of GCM test. In fact, it has been proved in Shah and Peters (2020) that GCM enjoys the so-called rate 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). It is not hard to show that GCM also enjoys the so-called model double-robustness property: if either $\widehat{\mu}_{n,x}(\cdot)$ or $\widehat{\mu}_{n,y}(\cdot)$ is consistently estimated at rate faster than $n^{-1/4}$, the test is valid. We formalize such model double-robustness property in the following theorem. We prove the result under regular iid setup, so we drop the subscript n for simplicity: $(X_{in}, Y_{in}, Z_{in}) = (X_i, Y_i, Z_i)$ and $\widehat{\mu}_{n,x} = \widehat{\mu}_x, \widehat{\mu}_{n,y} = \widehat{\mu}_y$.

Theorem 6 (Rate and model double-robustness of GCM)**.** *Define* $R_i \equiv (X_i - \widehat{\mu}_x(Z_i))(Y_i - \widehat{\mu}_y(Z_i))$ *and*

$$T_n^{\mathrm{GCM}}(X, Y, Z) \equiv \frac{1}{\sqrt{n}} \frac{\sum_{i=1}^n R_i}{\sqrt{\frac{1}{n}\sum_{i=1}^n R_i^2 - (\frac{1}{n}\sum_{i=1}^n R_i)^2}}.$$

Suppose there exist $C_u \in (0, \infty)$ *and functions* $\bar{\mu}_x, \bar{\mu}_y : \mathbb{R}^d \mapsto \mathbb{R}$ *(*$\bar{\mu}_x, \bar{\mu}_y$ *not necessarily equal to* μ_x, μ_y*) such that the following conditions hold.*

$$A_n \equiv \frac{1}{n}\sum_{i=1}^n \left(\widehat{\mu}_x(Z_i) - \bar{\mu}_x(Z_i)\right)^2 = o_{\mathbb{P}}(1); \quad (57)$$

$$B_n \equiv \frac{1}{n}\sum_{i=1}^n \left(\widehat{\mu}_y(Z_i) - \bar{\mu}_y(Z_i)\right)^2 = o_{\mathbb{P}}(1); \quad (58)$$

$$A_n \cdot B_n = o_{\mathbb{P}}(1/n); \quad (59)$$

$$\mathbb{E}[(\boldsymbol{X} - \bar{\mu}_x(\boldsymbol{Z}))^2 \mid \boldsymbol{Z}] \le C_u \quad \textit{and} \quad \mathbb{E}[(\boldsymbol{Y} - \bar{\mu}_y(\boldsymbol{Z}))^2 \mid \boldsymbol{Z}] \le C_u \ \textit{almost surely}; \quad (60)$$

$$\mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2 (Y_i - \bar{\mu}_y(Z_i))^2] > 0. \quad (61)$$

If either $\bar{\mu}_x = \mu_x$ *or* $\bar{\mu}_y = \mu_y$*, then under the null hypothesis* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$*, we have*

$$\lim_{n\to\infty} \mathbb{E}[\phi_{n,\alpha}^{\mathrm{GCM}}(X, Y, Z)] = \alpha \quad \textit{where} \quad \phi_{n,\alpha}^{\mathrm{GCM}}(X, Y, Z) \equiv \mathbb{1}\left(T_n^{\mathrm{GCM}}(X, Y, Z) > z_{1-\alpha}\right).$$

Proof of Theorem 6 can be found in Appendix D.3. Note that the result guarantees the validity of the test under the null hypothesis $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ as long as either $\mu_x(\cdot)$ or $\mu_y(\cdot)$ is consistently estimated at rate faster than $n^{-1/4}$. Then the test is valid even if the other estimator is misspecified.

The GCM test is related to doubly robust estimators in semiparametric inference, which are widely known to have the model double robustness property (see, e.g., Example 2 in Robins and Rotnitzky (2001)). Therefore, the model double-robustness of the GCM test is unsurprising and very much in line with these classical results.

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.

D.2 Illustrative comparison in the presence of sparsity

The important difference between dCRT and GCM 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. However, the asymptotic normal approximation may not be valid when there exists a large amount of sparsity in the data. In the next section, we argue from a theoretical perspective on how the rate of normal approximation can depend on the sparsity level of the data.

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 $\boldsymbol{X} \mid \boldsymbol{Z}$:

$$\boldsymbol{X} \mid \boldsymbol{Z} \sim \mathrm{Ber}(\mu_{n,x}(\boldsymbol{Z})). \tag{62}$$

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

$$T_n^{\mathrm{oGCM}}(X, Y, Z) \equiv \frac{1}{nS_n} \sum_{i=1}^{n} R_{in}^o \quad \text{where} \quad R_{in}^o \equiv (X_{in} - \mu_{n,x}(Z_{in}))(Y_{in} - \mu_{n,y}(Z_{in})), \tag{63}$$

and $S_n^2 = \mathbb{E}[(R_{in}^o)^2]$. Then we can define the test $\phi_{n,\alpha}^{\mathrm{oGCM}} \equiv \mathbb{1}(T_n^{\mathrm{oGCM}}(X, Y, Z) > z_{1-\alpha})$. For dCRT, we consider the modified dCRT with theoretical quantile, which we call oracle dCRT (odCRT):

$$\phi_{n,\alpha}^{\mathrm{odCRT}} \equiv \mathbb{1}\left(T_n^{\mathrm{dCRT}} \geq \mathbb{Q}_{1-\alpha}(\tilde{T}_n^{\mathrm{dCRT}} | X, Y, Z)\right) \quad \text{where} \quad \widetilde{X}^{(m)} \sim \prod_{i=1}^{n} \mathrm{Ber}(\mu_{n,x}(Z_{in})).$$

Here $T_n^{\mathrm{dCRT}}, \tilde{T}_n^{\mathrm{dCRT}}$ are defined in (14) and (15) respectively, but with oracle $\mu_{n,x}(\cdot), \mu_{n,y}(\cdot)$. The intuition for the Type-I error deviating from the specified significance level for GCM (and oGCM) is the CLT, when excessive sparsity exists, can happen in an arbitrarily slow rate depending on 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:* $\sup_z \mathbb{E}[\boldsymbol{Y}^4 \mid \boldsymbol{Z} = z] < \infty$, $\inf_z \mathbb{E}[(\boldsymbol{Y} - \mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{Z}])^3 \mid \boldsymbol{Z} = z] > 0$ *and* $\inf_z \mathrm{Var}[\boldsymbol{Y} \mid \boldsymbol{Z} = z] > 0$.

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

Parameter v_n in Assumption 6 characterizes the sparsity level of data, which will play an important role in Theorem 7 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 convergence on Type-I error for oGCM test. Such assumption can be satisfied 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 $\boldsymbol{Y} \mid \boldsymbol{Z}$. 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 (62) 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 7. *Consider* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. *Suppose Assumptions 6-8 hold. Then we have*

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

$$\left|\mathbb{E}[\phi_{n,\alpha}^{\text{oGCM}}] - \alpha - r_n\right| = o(r_n). \tag{64}$$

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

Remark 8 (Implication for testing with oGCM)**.** Theorem 7 unveils that sparsity in data can slow down the rate of Type-I error convergence to the specified significance level α. 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 7 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 that can be used to explain the finite-sample performance of dCRT and GCM tests.

D.3 Proof of Theorem 6

Proof of Theorem 6. We will prove the case when $\bar{\mu}_y = \mu_y$ since the other cases are similar. The proof consists of three steps.

1. We first show under the null hypothesis,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} R_i - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) = o_{\mathbb{P}}(1). \tag{65}$$

2. Then by Slutsky's theorem, it suffices to show that

$$\frac{1}{n} \sum_{i=1}^{n} R_i^2 - \mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2 (Y_i - \mu_y(Z_i))^2] = o_{\mathbb{P}}(1). \tag{66}$$

3. Last, we show that under the null hypothesis $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) \xrightarrow{d} N(0, \sigma^2)$$

where $\sigma^2 = \mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2 (Y_i - \mu_y(Z_i))^2]$.

The last step is true due to condition (61), $\mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2 (Y_i - \mu_y(Z_i))^2] \leq C_u^2 < \infty$ by condition (60), and an application of classical central limit theorem. Therefore, we just prove first two steps subsequently.

- **Proof of claim** (65)**.** To see this, we decompose

$$\begin{aligned}
\frac{1}{\sqrt{n}} R_i = & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{\mu}_x(Z_i))(\mu_y(Z_i) - \widehat{\mu}_y(Z_i)) \\
& + \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))(\mu_y(Z_i) - \widehat{\mu}_y(Z_i)) \\
\equiv & \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \bar{\mu}_x(Z_i))(Y_i - \mu_y(Z_i)) + \mathrm{Bias}_1 + \mathrm{Bias}_2 + \mathrm{Bias}_3.
\end{aligned}$$

Now we show that $\mathrm{Bias}_1, \mathrm{Bias}_2, \mathrm{Bias}_3$ are all $o_{\mathbb{P}}(1)$. We first show $\mathrm{Bias}_1 = o_{\mathbb{P}}(1)$. To see this, we first use conditoin (60) to bound

$$\begin{aligned}
\mathbb{E}[\mathrm{Bias}_1^2 | Y, Z] &= \frac{1}{n} \sum_{i=1}^{n} (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))^2 \mathrm{E}[(Y_i - \mu_y(Z_i))^2 | Z_i] \\
&\leq C_u \frac{1}{n} \sum_{i=1}^{n} (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))^2.
\end{aligned}$$

Then by conditonal Markov's inequality (Lemma 14), we have for any $\varepsilon > 0$ such that

$$\mathbb{P}[\mathrm{Bias}_1^2 > \varepsilon] = \mathbb{P}[\mathrm{Bias}_1^2 \wedge \varepsilon > \varepsilon] \leq \varepsilon^{-1} \mathbb{E}[\mathbb{E}[\mathrm{Bias}_1^2 | Y, Z] \wedge \varepsilon].$$

Then by condition (57) and dominated convergence theorem, we have $\mathbb{P}[\mathrm{Bias}_1^2 > \varepsilon] \to 0$. Thus we have shown that $\mathrm{Bias}_1 = o_{\mathbb{P}}(1)$. Similarly, we can show $\mathrm{Bias}_2 = o_{\mathbb{P}}(1)$ using condition (58). Now we show $\mathrm{Bias}_3 = o_{\mathbb{P}}(1)$. To see this, we use Cauchy-Schwarz inequality to compute

$$\mathrm{Bias}_3 \leq \sqrt{n} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^{n} (\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2}.$$

By Assumption (59), we have $\mathrm{Bias}_3 = o_{\mathbb{P}}(1)$. Thus we have shown that $\mathrm{Bias}_1, \mathrm{Bias}_2, \mathrm{Bias}_3 = o_{\mathbb{P}}(1)$ and hence (65) holds.

- **Proof of claim (66).** It suffices to show $\frac{1}{n}\sum_{i=1}^n R_i^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2] > 0$. To see this, we decompose

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n R_i^2 = \frac{1}{n}\sum_{i=1}^n &(X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2 \\
&+ \frac{1}{n}\sum_{i=1}^n (X_i - \bar{\mu}_x(Z_i))^2(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 \\
&+ \frac{2}{n}\sum_{i=1}^n (X_i - \bar{\mu}_x(Z_i))^2(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))(Y_i - \mu_y(Z_i)) \\
&+ \frac{1}{n}\sum_{i=1}^n (\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))^2(Y_i - \widehat{\mu}_y(Z_i))^2 \\
&+ \frac{2}{n}\sum_{i=1}^n (X_i - \bar{\mu}_x(Z_i))(\bar{\mu}_x(Z_i) - \widehat{\mu}_x(Z_i))(Y_i - \widehat{\mu}_y(Z_i))^2 \\
\equiv \frac{1}{n}\sum_{i=1}^n &(X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2 + \sum_{j=1}^4 C_j.
\end{aligned}$$

It suffices to show that C_1, C_2, C_3, C_4 are all $o_{\mathbb{P}}(1)$ and conclude the proof by using weak law of large numbers. We will just show C_1, C_2 are $o_{\mathbb{P}}(1)$ and the other two terms are similar. To see this is true for C_1, we first use conditoin (60) to bound

$$\begin{aligned}
\mathbb{E}[C_1|Y, Z] &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2|Z](\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2 \\
&\leq C_u \frac{1}{n}\sum_{i=1}^n (\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^2.
\end{aligned}$$

Then by conditional Markov's inequality (Lemma 14), we have for any $\varepsilon > 0$ such that

$$\mathbb{P}[C_1 > \varepsilon] = \mathbb{P}[C_1 \wedge \varepsilon > \varepsilon] \leq \varepsilon^{-1}\mathbb{E}[\mathbb{E}[C_1|Y, Z] \wedge \varepsilon].$$

Then by condition (58) and dominated convergence theorem, we have $\mathbb{P}[C_1 > \varepsilon] \to 0$. Thus we have shown that $C_1 = o_{\mathbb{P}}(1)$. Now we show C_2 is $o_{\mathbb{P}}(1)$. To see this, we first apply Cauchy-Schwarz inequality to bound

$$|C_2| \leq 2\sqrt{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2} \cdot \sqrt{C_1}.$$

Then by law of large numbers, we have $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2 \xrightarrow{\mathbb{P}} \mathbb{E}[(X_i - \bar{\mu}_x(Z_i))^2(Y_i - \mu_y(Z_i))^2]$. Then since we have proved $C_1 = o_{\mathbb{P}}(1)$ we conclude the proof of $C_2 = o_{\mathbb{P}}(1)$. The other two terms can be proved similarly. Thus we have shown that $C_1, C_2, C_3, C_4 = o_{\mathbb{P}}(1)$ and hence (66) holds.

□

D.4 Proof of Theorem 7

To prove Theorem 7, we first need an auxiliary result.

Lemma 6 (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 $x \in \mathbb{R}$,*

$$\left| \mathbb{P}\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} W_{in} \leq x \right] - \mathbb{P}[Z \leq x] - \frac{(1-x^2)\exp(-x^2/2)\mathbb{E}[W_{in}^3]}{6\sqrt{2\pi n}} \right| \leq C\frac{\mathbb{E}[W_{in}^4]}{n} + C\left(\sup_{|t| \geq 1/(12\mathbb{E}[|W_{in}|^3])} |\mathbb{E}[\exp(itW_{in})]| + \frac{1}{2n} \right)^n \frac{n^6}{1+|x|^4}. \tag{67}$$

Proof of Theorem 7. The argument for the validity of odCRT is based on the exchangeability of the resampled data and the original data under null hypothesis $H_0 : \boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$. Now we prove the convergence rate of oGCM. In order to prove the result, we will use Lemma 6 to state an asymptotic expansion of CDF of oGCM test statistic (63). We apply $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ in Lemma 6 so that we get a bound as in (64) with the desired 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 the RHS of (67) with $W_{in} = S_{in}$ and $x = z_{1-\alpha}$ is of smaller order of r_n. In fact, 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 convergence rate of r_n.

Convergence rate of r_n. By conditional independence $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$, it is easy to show that

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

Then by Hölder's inequality, we know

$$\mathbb{E}[(Y_{in}-\mathbb{E}[Y_{in}|Z_{in}])^3|Z_{in}=z] \lesssim \mathbb{E}[Y_{in}^4|Z_{in}=z] \leq \sup_z \mathbb{E}[Y_{in}^4|Z_{in}=z] < \infty.$$

Then by Assumption 7, we have $r_n \asymp 1/(n^{1/2}v_n^{1/2})$.

Concergence rate of the term involving $|\mathbb{E}[\exp(itS_{in})]|$. 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

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

so that we prove

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

Convergence rate of the term involving $\mathbb{E}[S_{in}^4]$. It remains to prove $\frac{\mathbb{E}[S_{in}^4]}{nr_n} = o(1)$. It suffices to show $\mathbb{E}[S_{in}^4] = o(n^{1/2}/v_n^{1/2})$ by the proved results $r_n \asymp 1/(nv_n)^{1/2}$. To see this, by Assumption 7, we can bound

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

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

E Additional details of Section 4

E.1 Asymptotic equivalence of dCRT and spaCRT

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

$$(\widehat{S}_n^{\mathrm{dCRT}})^2 \equiv \frac{1}{n} \sum_{i=1}^n \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in} \mid Z_{in}](Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2. \tag{68}$$

Also define the asymptotic test:

$$\phi_{n,\alpha}^{\mathrm{asy}} \equiv \mathbb{1}\left(\frac{n^{1/2} T_n^{\mathrm{dCRT}}(X,Y,Z)}{\widehat{S}_n^{\mathrm{dCRT}}} > z_{1-\alpha} \right). \tag{69}$$

Theorem 8 (Asymptotic equivalence of tests)**.** *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* (68)*, does not accumulate around the* $1-\alpha$ *quantile of standard normal distribution* $z_{1-\alpha}$*, i.e.,*

$$\lim_{\delta\to 0}\limsup_{n\to\infty}\mathbb{P}_{\mathcal{L}_n}\left[\left|\frac{n^{1/2}T_n^{\mathrm{dCRT}}(X,Y,Z)}{\widehat{S}_n^{\mathrm{dCRT}}}-z_{1-\alpha}\right|\le\delta\right]=0, \tag{70}$$

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 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 7 (Bernoulli sampling)**.** *Suppose* $\boldsymbol{X}$ *is a binary variable following the natural exponential family model* (13) *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:

- ***Condition set 1:***

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

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

- ***Condition set 2:***

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

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

- ***Condition set 3:***

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

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

Then if $T_n^{\mathrm{dCRT}}(X,Y,Z)\xrightarrow{\mathbb{P}}0$*, the conclusion in Theorem 2 holds and*

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

Lemma 7 will be used to prove Theorem 3, Theorem 4 and Theorem 9 (present in next section).

E.3 Application of spaCRT to kernel ridge regression

In this section, we study the validity of spaCRT when the conditional distribution $\boldsymbol{Y} \mid \boldsymbol{Z}$ 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 $(X_{in}, Y_{in}, Z_{in}) = (X_i, Y_i, Z_i)$ and $\gamma_n = \gamma$.

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

$$k(z, z') = \sum_{j=1}^{\infty} \kappa_j e_j(z) e_j(z') \tag{79}$$

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 (Theorem 12.20, Wainwright, 2019) if mild conditions are satisfied. For a sequence of regularization parameter λ_n, we consider the following estimator:

$$\widehat{\mu}_y \equiv \arg\min_{\mu_y \in \mathcal{H}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_y(Z_i))^2 + \lambda_n \|\mu_y\|_{\mathcal{H}}^2 \right\}. \tag{80}$$

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

$$\lambda_n = \arg\min_{\lambda > 0} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{\kappa}_i^2}{(\widehat{\kappa}_i + \lambda)^2} + \lambda \right\} \tag{81}$$

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 $\widehat{\gamma}$ for γ, we consider using the maximum likelihood estimator $\widehat{\gamma}$. With the estimators $\widehat{\mu}_y(Z_i)$ and $\widehat{\theta}_{n,x}(Z_i) = Z_i^\top \widehat{\gamma}$, the spaCRT can be applied with Algorithm 2. Now we state our main results on the validity guarantee of spaCRT.

Theorem 9. *Suppose* $\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y} \mid \boldsymbol{Z}$ *and that Assumptions 3-4 hold. Then if the following conditions hold:*

$$\text{support of } \boldsymbol{Y} \text{ is compact, i.e., there exists } S, \mathbb{P}[\boldsymbol{Y} \in [-S, S]] = 1; \tag{82}$$

$$\|\widehat{\gamma} - \gamma\|_1 = O_{\mathbb{P}}(1/\sqrt{n}); \tag{83}$$

$$\|\widehat{\mu}_y\|_\infty = O_{\mathbb{P}}(1); \tag{84}$$

$$\sum_{j=1}^{\infty} \kappa_j < \infty, \tag{85}$$

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

Conditions in Theorem 9 are mild conditions. Condition (84) can be easily verified when linear kernel, i.e. linear ridge regression, is considered. For general choice of kernel, condition (84) can hold under extra conditions on the kernel function k. The proof is postponed to Section O.

E.3.1 Simulation on unbalanced nonparametric classification

Simulation setup. We consider a nonlinear classification problem with unbalanced class sizes. We consider the following data generating procedure. First generate $\boldsymbol{Z}_j$ independent from the uniform distribution on $[-1, 1]$ and let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)^\top$ be the covariate vector. The response $\boldsymbol{X}$ and $\boldsymbol{Y}$ is generated from the following model:

$$\boldsymbol{X}|\boldsymbol{Z} \sim \text{Ber}(\text{expit}(\gamma_0 + g(\boldsymbol{Z}))) \quad \text{and} \quad \boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{Z} \sim \text{Ber}(\text{expit}(\gamma_0 + \boldsymbol{X}\eta + g(\boldsymbol{Z})))$$

where $g(\boldsymbol{Z}) = \sin(\pi \boldsymbol{Z}_1) + \boldsymbol{Z}_2^2 \boldsymbol{Z}_3$. Intercept γ_0 controls the sparsity of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and η controls signal strength. We set $\gamma_0 \in \{-2.5, -2\}$; we set η to vary within $\{0, 1, 2, 3, 4\}$ when $\gamma_0 = -2.5$ and $\{0, 0.5, 1, 1.5, 2\}$ when $\gamma_0 = -2$.

Methodologies compared. We consider three tests: GCM test, dCRT and spaCRT. We use `probability_forest` from R package `grf`, a variant of random forest for classification, to fit the model $\boldsymbol{X}|\boldsymbol{Z}$ and $\boldsymbol{Y}|\boldsymbol{Z}$. We use the default hyperparameters in the function. We set the number of resample for dCRT to be $M = 50000$. The significance level is set to be $\alpha = 0.005$.

Simulation results. The results on Type-I error control, power and computation time are summarized in Figure 6. We can see that spaCRT and dCRT have similar power and Type-I error control, while GCM test suffers from inflated Type-I error. spaCRT is similar to GCM test in terms of computational time but 2 times faster than dCRT. We found the failure to solve the saddlepoint equation quite rare, occurring in at most 0.002% of replications across all simulation settings.

F Some useful lemmas and proofs

F.1 Lemma statements

Lemma 8 (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:

$$\lim_{\delta \to 0} \limsup_{n \to \infty} \mathbb{P}_{\mathcal{L}_n}[|T_n(X, Y, Z) - z_{1-\alpha}| \leq \delta] = 0, \tag{86}$$

then the two tests are asymptotically equivalent:

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

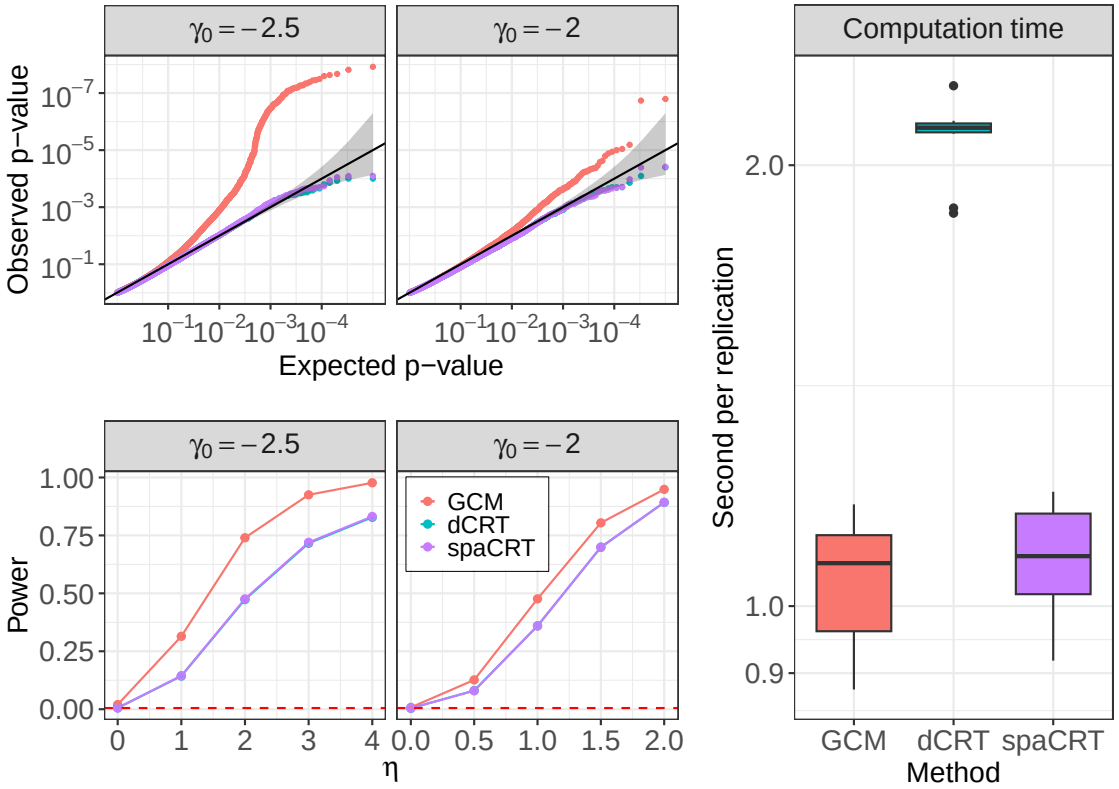

Figure 6: Summary of numerical simulation results for nonparametric regression. The simulation is repated 50000 times.

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:

$$(\widehat{S}_n^{\mathrm{dCRT}})^2 = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in} \mid Z_{in}](Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2 = \Omega_{\mathbb{P}}(1); \tag{87}$$

$$\frac{1}{n^{1+\delta/2}}\sum_{i=1}^{n} |Y_{in} - \widehat{\mu}_{n,y}(Z_{in})|^{2+\delta}\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in})|^{2+\delta} \mid X, Z] = o_{\mathbb{P}}(1); \tag{88}$$

$$\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in}|Z_{in}], (Y_{in} - \widehat{\mu}_{n,y}(Z_{in}))^2, (Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty \text{ almost surely.} \tag{89}$$

Lemma 9 (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*

$$T_n^{\mathrm{ndCRT}}(\widetilde{X}, X, Y, Z) \equiv \frac{T_n^{\mathrm{dCRT}}(\widetilde{X}, X, Y, Z)}{\widehat{S}_n^{\mathrm{dCRT}}} \tag{90}$$

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 10. *Suppose the sequence of laws $\mathcal{L}_n$ and its estimate $\widehat{\mathcal{L}}_n$ satisfy all the assumptions in Lemma 9. Then for any given $\alpha \in (0,1)$, we have for any sequence $M_n \in \mathcal{F}_n$ satisfying $M_n = o_{\mathbb{P}}(1)$,*

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

where $T_n^{\mathrm{ndCRT}}(\widetilde{X}, X, Y, Z)$ is defined as in (90).

Lemma 11 (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*

$$\frac{1}{n^{1+\delta}} \sum_{i=1}^{n} \mathbb{E}[|X_{in}|^{1+\delta}] \to 0, \tag{91}$$

then

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

The condition (91) is satisfied when

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

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

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

Lemma 13 (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*

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

Lemma 14 (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*

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

Lemma 15 (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 16. *Consider the fixed dimension setup:* $(X_{in}, Y_{in}, Z_{in}) = (X_i, Y_i, Z_i)$ *and* $(Z_i, X_i, Y_i)_{i\in[n]}$ *are i.i.d. samples. 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* $\boldsymbol{Y} \mid \boldsymbol{Z} \sim f(\boldsymbol{Y} \mid \boldsymbol{Z}^\top\beta)$ *for some natural exponential family* f (13) *with log-partition function* A_y *and* $\mathbb{P}[\|Z_i\|_\infty, X_i, \mu_x(Z_i) \in [-S, S]] = 1$ *for some* $S > 0$, *then as long as the following conditions hold:*

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

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

Lemma 17 (Gaussian tail probability estimate)**.** *For* $x > 0$, *we have*

$$1 - \Phi(x) - \frac{1}{\sqrt{2\pi}x}\exp(-x^2/2)\left(1 - \frac{1}{x^2}\right) = -\int_x^\infty \frac{\phi(t)}{3t^4}\mathrm{d}t.$$

Consequently, we have

$$\left|1 - \Phi(x) - \left(\frac{1}{\sqrt{2\pi}x}\exp(-x^2/2)\left(1 - \frac{1}{x^2}\right)\right)\right| \le \frac{\phi(x)}{x^3}$$

and

$$\left|x\exp\left(\frac{x^2}{2}\right)(1 - \Phi(x)) - \frac{1}{\sqrt{2\pi}}\right| \le \frac{2}{\sqrt{2\pi}}\frac{1}{x^2}.$$

Lemma 18 (Lower bound on the Gaussian tail probability)**.** *For any* $x \ge 0$, *we have*

$$1 - \Phi(x) > \frac{1}{\sqrt{2\pi}}\frac{x}{x^2+1}\exp(-x^2/2).$$

Lemma 19. *Consider the probability space* $(\mathbb{P}, \Omega, \mathcal{F})$ *and the* σ*-algebras* $\mathcal{F}_n \subset \mathcal{F}$. *Suppose the sequence of random variable* W_n *satisfies there exists* $\varepsilon > 0$ *such that*

$$\mathbb{P}\left[\mathbb{E}[|W_n|^p\exp(sW_n)|\mathcal{F}_n] < \infty,\ \forall s \in (-\varepsilon, \varepsilon),\ \forall n, p \in \mathbb{N}\right] = 1$$

Then defining $H_n(s) \equiv \mathbb{E}[\exp(sW_n)|\mathcal{F}_n]$, *we have*

$$\mathbb{P}\left[H_n(s) \text{ has } p\text{-th order derivative at the open neighborhood } (-\varepsilon, \varepsilon),\ \forall n, p \in \mathbb{N}\right] = 1,$$

and

$$\mathbb{P}\left[H_n^{(p)}(s) = \mathbb{E}[W_n^p\exp(sW_n)|\mathcal{F}_n], \forall s \in (-\varepsilon, \varepsilon),\ \forall n, p \in \mathbb{N}\right] = 1.$$

Lemma 20 (Chen, Goldstein, and Shao, 2011, Theorem 3.6)**.** *Suppose $n \in \mathbb{N}$ and $\xi_{1n}, \ldots, \xi_{nn}$ are independent random variables, satisfying for any $1 \leq i \leq n$*

$$\mathbb{E}[\xi_{in}] = 0, \ \sum_{i=1}^{n} \mathbb{E}[\xi_{in}^2] = 1.$$

Then

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left[\sum_{i=1}^{n} \xi_{in} \leq t \right] - \Phi(t) \right| \leq 9.4 \sum_{i=1}^{n} \mathbb{E}[|\xi_{in}|^3].$$

Lemma 21 (Equivalence of the definition of CSE distribution)**.** *The following two statements are equivalent:*

1. *there exists positive parameters (λ_{in}, γ) with $\lambda_{in} \in \mathcal{F}_n$ and constant γ such that*

$$\mathbb{P}\left[\mathcal{B}_1\right] = 1, \ \mathcal{B}_1 \equiv \left\{ \mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] \leq \exp(\lambda_{in} s^2), \ \forall s \in \left(-\frac{1}{\gamma}, \frac{1}{\gamma} \right) \right\}. \tag{92}$$

2. *there exists positive parameters (θ_{in}, β) with $\theta_{in} \in \mathcal{F}_n$ and constant β such that*

$$\mathbb{P}\left[\mathcal{B}_2\right] = 1, \ \mathcal{B}_2 \equiv \left\{ \mathbb{P}\left[|W_{in}| \geq t|\mathcal{F}_n\right] \leq \theta_{in} \exp(-\beta t), \ \forall t > 0 \right\}. \tag{93}$$

In particular, the suppose condition (93) holds, then we can choose (λ_{in}, γ) in (92) as

$$\lambda_{in} = \frac{\sqrt{6!4^6}(1+\theta_{in})}{24\beta^2} + \frac{16(1+\theta_{in})}{\beta^2}, \ \gamma = \frac{4}{\beta}.$$

F.2 Proof of Lemma 10

Proof of Lemma 10. For any given $\varepsilon \in (0, \min\{1/\alpha - 1, 1\}), \eta > 0$, there exists $N(\varepsilon, \eta)$ such that

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

This is true because $M_n = o_{\mathbb{P}}(1)$. 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 $1 - \alpha(1+\varepsilon)$ conditional quantiles of T_n^{ndCRT}. We have with probability at least $1 - \eta$, for large enough n, the following is true:

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

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

$$\left|\mathbb{Q}_{1-\alpha(1+M_n)}\left[T_n^{\text{ndCRT}}|X, Y, Z\right] - z_{1-\alpha}\right| \leq A_n + B_n \tag{94}$$

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 9, we have

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

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 (94), 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

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

Therefore we complete the proof. □

F.3 Proof of Lemma 15

Proof of Lemma 15. By Lemma 12, we have

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

We use Jensen's inequality, Lemma 13, to obtain

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

□

F.4 Proof of Lemma 16

Proof of Lemma 16. We consider the following decomposition:

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

We will prove $\sqrt{n}B_k$ converges 0 in probability for any $k = 1, 2, 3$. First, by the assumption $Y_i \sim f(\cdot \mid Z_i^\top \beta)$ for NEF f with log-partition function A_y and Hölder's inequality $|Z_i^\top \beta| \le \|Z_i\|_\infty \|\beta\|_1 \le S\|\beta\|_1$, we have, almost surely,

$$\mathbb{E}[(Y_i - \mu_y(Z_i))^2 | Z] = A_y''(Z_i^\top \beta) \le \sup_{t \in [-S\|\beta\|_1, S\|\beta\|_1]} A''(t) < \infty.$$

Therefore, for B_1, we have

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

For B_2, by the assumption that $\mathbb{P}[X_i, \mu_x(Z_i) \in [-S, S]] = 1$, we have

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

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 classical CLT, guaranteed by the assumption that $\sigma_{\text{dCRT}} \in (0, \infty)$. □

F.5 Proof of the Lemma 17

Proof of Lemma 17. Applying integration by parts, we can write

$$\begin{aligned}
1 - \Phi(x) &= \int_x^\infty \phi(t) \mathrm{d}t \\
&= \int_x^\infty \frac{1}{t} \frac{t}{\sqrt{2\pi}} \exp(-t^2/2) \mathrm{d}t \\
&= -\frac{1}{t} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \Big|_x^\infty - \int_x^\infty \frac{\phi(t)}{t^2} \mathrm{d}t \\
&= \frac{\phi(x)}{x} + \frac{1}{t^3} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) \Big|_x^\infty - \int_x^\infty \frac{\phi(t)}{3t^4} \mathrm{d}t \\
&= \frac{\phi(x)}{x} - \frac{\phi(x)}{x^3} - \int_x^\infty \frac{\phi(t)}{3t^4} \mathrm{d}t.
\end{aligned}$$

Then we can bound for $x > 0$

$$\left| \int_x^\infty \frac{\phi(t)}{3t^4} \mathrm{d}t \right| \le \phi(x) \int_x^\infty \frac{1}{3t^4} \mathrm{d}t \le \frac{\phi(x)}{x^3}.$$

□

F.6 Proof of Lemma 18

Proof of Lemma 18. Define

$$g(x) \equiv 1 - \Phi(x) - \frac{1}{\sqrt{2\pi}} \frac{x}{x^2+1} \exp(-x^2/2).$$

Computing the derivative we obtain

$$g'(x) = -\frac{2}{\sqrt{2\pi}} \frac{e^{-x^2/2}}{(x^2+1)^2} < 0.$$

Also notice $g(0) = 1/2 > 0$ and $\lim_{x\to\infty} g(x) = 0$. This completes the proof. □

F.7 Proof of Lemma 19

Proof of Lemma 19. Consider the regular conditional distribution $W_n|\mathcal{F}_n$ to be κ_{W_n}. We use induction to prove the existence of the p-th derivative of $H_n(s)$. Suppose

$$\mathbb{P}\left[H_n^{(p)}(s) = \mathbb{E}[W_n^p \exp(sW_n)|\mathcal{F}_n],\ \forall s \in (-\varepsilon, \varepsilon)\right] = 1 \tag{95}$$

and

$$\mathbb{P}\left[\mathbb{E}[|W_n|^{p+1} \exp(sW_n)|\mathcal{F}_n] < \infty,\ \forall s \in (-\varepsilon, \varepsilon)\right] = 1. \tag{96}$$

According to the definition of derivative, we write

$$H_n^{(p+1)}(s) \equiv \lim_{h\to 0} \frac{H_n^{(p)}(s+h) - H_n^{(p)}(s)}{h}. \tag{97}$$

Then on the event in hypothesis (95), we have for any $s \in (-\varepsilon, \varepsilon)$

$$H_n^{(p)}(s) = \mathbb{E}[W_n^p \exp(sW_n)|\mathcal{F}_n] = \int x^p \exp(sx) \mathrm{d}\kappa_{W_n}(\cdot, x).$$

Fix any $s_0 \in (-\varepsilon, \varepsilon)$ and find $r_0 \in (s_0, \varepsilon)$ such that $|r_0| > |s_0|$. We find small enough h such that $|h| < \min\{|r_0| - s_0, |r_0| + s_0\}$. Thus we have

$$-\varepsilon < -|r_0| = s_0 - |r_0| - s_0 < s_0 + h < s_0 + |r_0| - s_0 = |r_0| < \varepsilon. \tag{98}$$

Also we notice, $|s_0| < |r_0|$ so that $s_0 x \in (-|r_0 x|, |r_0 x|)$. Also, the derivation in (98) informs $(s_0+h)x$ belong to the interval $(-|r_0 x|, |r_0 x|)$. Therefore, both $s_0 x$ and $(s_0+h)x$ belong to $(-|r_0 x|, |r_0 x|)$. Then we have

$$\begin{aligned} |\exp((s_0+h)x) - \exp(s_0 x)| &= \left| \int_{s_0 x}^{(s_0+h)x} e^y \mathrm{d}y \right| \\ &\le |hx| \sup_{y \in [-|r_0 x|, |r_0 x|]} \exp(y) \\ &\le |hx| \{\exp(r_0 x) + \exp(-r_0 x)\}. \end{aligned} \tag{99}$$

By the definition (97), we have

$$H_n^{(p+1)}(s_0) = \lim_{h\to 0} \mathbb{E}\left[W_n^p \frac{\exp((s_0+h)W_n) - \exp(s_0 W_n)}{h} | \mathcal{F}_n\right]$$
$$= \lim_{h\to 0} \int x^p \frac{\exp((s_0+h)x) - \exp(s_0 x)}{h} \mathrm{d}\kappa_{W_n}(\cdot, x).$$

Then by (99), we can bound

$$\left| x^p \frac{\exp((s_0+h)x) - \exp(s_0 x)}{h} \right| \leq |x|^{p+1} \exp(r_0 x) + |x|^{p+1} \exp(-r_0 x).$$

Notice the RHS is independent of h and integrable with respect to measure $\kappa_{W_n}(\omega, \cdot)$ for almost every $\omega \in \Omega$, by the induction hypothesis (96) since $r_0 \in (-\varepsilon, \varepsilon)$. By dominated convergence theorem, we know on the event in hypothesis (96),

$$H_n^{(p+1)}(s_0) = \int \lim_{h\to 0} x^p \frac{\exp((s_0+h)x) - \exp(s_0 x)}{h} \mathrm{d}\kappa_{W_n}(\cdot, x)$$
$$= \int x^{p+1} \exp(s_0 x) \mathrm{d}\kappa_{W_n}(\cdot, x)$$
$$= \mathbb{E}[W_n^{p+1} \exp(s_0 W_n) | \mathcal{F}_n].$$

Then, by the arbitrary choice of $s_0 \in (-\varepsilon, \varepsilon)$, we know on the event in hypothesis (96), $H^{(p+1)}(s)$ is well-defined on the interval $(-\varepsilon, \varepsilon)$ and takes the form

$$H^{(p+1)}(s) = \mathbb{E}[W_n^{p+1} \exp(s_0 W_n) | \mathcal{F}_n].$$

Thus we have proved for the case $p+1$ so that we complete the induction and conclude the proof. □

F.8 Proof of Lemma 21

Proof of Lemma 21. We prove two directions separately.

(92)⇒(93): For any $\omega \in \mathcal{B}_1$, we know by definition (32) that

$$\mathbb{E}[\exp(sW_{in}) | \mathcal{F}_n](\omega) = \int \exp(sx) \mathrm{d}\kappa_{in}(\omega, x) < \infty, \ \forall s \in \left(-\frac{1}{\gamma}, \frac{1}{\gamma}\right).$$

Then the Chernoff bound gives

$$\int \mathbb{1}(x \geq t) \mathrm{d}\kappa_{in}(\omega, x) \leq \int \exp\left(\frac{x}{2\gamma}\right) \mathrm{d}\kappa_{in}(\omega, x) \exp\left(-\frac{t}{2\gamma}\right).$$

Applying a similar argument to $\mathbb{1}(-x \geq t)$, we conclude

$$\int \mathbb{1}(|x| \geq t) \mathrm{d}\kappa_{in}(\omega, x) \leq \left(\int \exp\left(\frac{x}{2\gamma}\right) + \exp\left(\frac{-x}{2\gamma}\right) \mathrm{d}\kappa_{in}(\omega, x)\right) \cdot \exp\left(-\frac{t}{2\gamma}\right).$$

Thus we know

$$\begin{aligned}\mathbb{P}\left[|W_{in}| \geq t|\mathcal{F}_n\right] &= \int \mathbb{1}(|x| \geq t)\mathrm{d}\kappa_{in}(\omega, x) \\ &\leq \left(\mathbb{E}\left[\exp(W_{in}/(2\gamma))|\mathcal{F}_n\right] + \mathbb{E}\left[\exp(-W_{in}/(2\gamma))|\mathcal{F}_n\right]\right) \cdot \exp(-t/(2\gamma)).\end{aligned}$$

Thus $\mathbb{P}[\mathcal{B}_2] = 1$ with

$$\theta_{in} = \mathbb{E}\left[\exp(W_{in}/(2\gamma))|\mathcal{F}_n\right] + \mathbb{E}\left[\exp(-W_{in}/(2\gamma))|\mathcal{F}_n\right], \ \beta = \frac{1}{2\gamma}.$$

(93)⇒(92): Fix a constant $a > 0$ and $T > 0$. For any $\omega \in \mathcal{B}_2$, we have

$$\begin{aligned}&\mathbb{E}\left[\exp(a|W_{in}|)\mathbb{1}(\exp(a|W_{in}|) \leq \exp(aT))|\mathcal{F}_n\right](\omega) \\ &= \int \exp(a|x|)\mathbb{1}(\exp(a|x|) \leq \exp(aT))\mathrm{d}\kappa_{in}(\omega, x) \\ &\leq \int \min\{\exp(a|x|), \exp(aT)\}\mathrm{d}\kappa_{in}(\omega, x) \\ &= \int \int_0^{e^{aT}} \mathbb{1}(\exp(a|x|) \geq t)\mathrm{d}t\mathrm{d}\kappa_{in}(\omega, x) \\ &= \int_0^{e^{aT}} \int \mathbb{1}(\exp(a|x|) \geq t)\mathrm{d}\kappa_{in}(\omega, x)\mathrm{d}t \\ &\leq 1 + \int_1^{e^{aT}} \int \mathbb{1}(|x| \geq \log(t)/a)\mathrm{d}\kappa_{in}(\omega, x)\mathrm{d}t\end{aligned}$$

where the last equality is due to Fubini's theorem. Then by the definition of $\mathcal{B}_2$, we obtain

$$\begin{aligned}\mathbb{E}\left[\exp(a|W_{in}|)\mathbb{1}(\exp(a|W_{in}|) \leq \exp(aT))|\mathcal{F}_n\right](\omega) &\leq 1 + \theta_{in}(\omega)\int_1^{e^{aT}} e^{-(\beta\log(t))/a}\mathrm{d}t \\ &= 1 + \theta_{in}(\omega)\int_1^{e^{aT}} t^{-\beta/a}\mathrm{d}t.\end{aligned}$$

For $a \in [0, \beta/2]$, we have

$$\begin{aligned}\mathbb{E}\left[\exp(a|W_{in}|)\mathbb{1}(\exp(a|W_{in}|) \leq \exp(aT))|\mathcal{F}_n\right](\omega) &= 1 + \theta_{in}(\omega)\frac{1}{1-\beta/a}t^{1-\beta/a}\Big|_1^{e^{aT}} \\ &\leq 1 + \theta_{in}(\omega)\frac{1}{\beta/a - 1}(1 - e^{(aT-\beta T)}) \\ &\leq 1 + \theta_{in}(\omega).\end{aligned}$$

Then by Fatou's lemma, we have for any $a \in [0, \beta/2]$,

$$\begin{aligned}\mathbb{E}\left[\exp(a|W_{in}|)|\mathcal{F}_n\right](\omega) &= \int \exp(a|x|)\mathrm{d}\kappa_{in}(\omega, x) \\ &\leq \liminf_{T\to\infty} \int \exp(a|x|)\mathbb{1}(e^{a|x|} \leq e^{aT})\mathrm{d}\kappa_{in}(\omega, x) \leq 1 + \theta_{in}(\omega).\end{aligned} \tag{100}$$

Then by Taylor's expansion, for any $|s| \leq \beta/4$

$$\begin{aligned}\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n](\omega) &= \int \exp(sx)\mathrm{d}\kappa_{in}(\omega, x) \\ &= \int 1 + sx + \frac{s^2x^2}{2} + \frac{s^3x^3}{6}\exp(y(x))\mathrm{d}\kappa_{in}(\omega, x),\ |y(x)| \leq |sx|.\end{aligned}$$

Then the assumption $\mathbb{E}[W_{in}|\mathcal{F}_n] = 0$ implies for $|s| \leq \beta/4$,

$$\begin{aligned}\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n](\omega) &\leq 1 + \frac{s^2\mathbb{E}[W_{in}^2|\mathcal{F}_n](\omega)}{2} + \frac{s^2\beta}{24}\int |x|^3\exp(y(x))\mathrm{d}\kappa_{in}(\omega, x) \\ &\leq 1 + \frac{s^2\mathbb{E}[W_{in}^2|\mathcal{F}_n](\omega)}{2} + \frac{s^2\beta}{24}\int |x|^3\exp(|sx|)\mathrm{d}\kappa_{in}(\omega, x) \\ &\leq 1 + \frac{s^2\mathbb{E}[W_{in}^2|\mathcal{F}_n](\omega)}{2} + \frac{s^2\beta}{24}\sqrt{\mathbb{E}[W_{in}^6|\mathcal{F}_n](\omega)\mathbb{E}[\exp(2|sW_{in}|)|\mathcal{F}_n](\omega)} \\ &\leq 1 + \left(\frac{\beta\sqrt{1+\theta_{in}(\omega)}\sqrt{\mathbb{E}[W_{in}^6|\mathcal{F}_n](\omega)}}{12} + \mathbb{E}[W_{in}^2|\mathcal{F}_n](\omega)\right)\frac{s^2}{2} \\ &\leq \exp\left(\left(\frac{\beta\sqrt{1+\theta_{in}(\omega)}\sqrt{\mathbb{E}[W_{in}^6|\mathcal{F}_n](\omega)}}{12} + \mathbb{E}[W_{in}^2|\mathcal{F}_n](\omega)\right)\frac{s^2}{2}\right).\end{aligned} \tag{101}$$

where the second inequality is due to $|y(x)| \leq |sx|$, the third inequality is due to Cauchy-Schwarz inequality, the fourth inequality is due to conclusion (100) and the last inequality is due to the inequality $\exp(x) \geq 1 + x$ for any $x \in \mathbb{R}$. Now by Fubini's theorem and conclusion (100), we have

$$1 + \theta_{in}(w) \geq \mathbb{E}[\exp(s|W_{in}|)|\mathcal{F}_n] = 1 + \sum_{k=1}^{\infty}\frac{s^k}{k!}\mathbb{E}[|W_{in}|^k|\mathcal{F}_n],\ \forall s \in [0, \beta/2].$$

Then by setting $s = \beta/4$, we have

$$\begin{aligned}\mathbb{E}[W_{in}^2|\mathcal{F}_n] &\leq \frac{2!4^2}{\beta^2}\mathbb{E}[\exp(\beta|W_{in}|/4)|\mathcal{F}_n] \leq \frac{2!4^2}{\beta^2}(1+\theta_{in}), \\ \mathbb{E}[W_{in}^6|\mathcal{F}_n] &\leq \frac{6!4^6}{\beta^6}\mathbb{E}[\exp(\beta|W_{in}|/4)|\mathcal{F}_n] \leq \frac{6!4^6}{\beta^6}(1+\theta_{in})\end{aligned}$$

Then choosing

$$\lambda_{in} = \frac{\sqrt{6!4^6}(1+\theta_{in})}{24\beta^2} + \frac{16(1+\theta_{in})}{\beta^2} \geq \frac{\beta\sqrt{1+\theta_{in}}\sqrt{\mathbb{E}[W_{in}^6|\mathcal{F}_n]}}{24} + \frac{\mathbb{E}[W_{in}^2|\mathcal{F}_n]}{2},\ \gamma = \frac{4}{\beta},$$

so that by bound (101), we obtain

$$\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n](\omega) \leq \exp(\lambda_{in}(\omega)s^2),\ \forall s \in \left(-\frac{1}{\gamma}, \frac{1}{\gamma}\right).$$

□

G Proof of Theorem 1

The high-level structure of our proof is inspired by that of Robinson (1982): Exponentially tilt the summands, then apply the Berry-Esseen inequality to get a normal approximation after tilting, then tilt back. In this section, we sketch the proof our main result with the help of a sequence of lemmas, whose proofs we defer to Appendix I.

G.1 Solving the saddlepoint equation

First, we state a lemma lower-bounding the second derivative $K_n''(s)$, which will help us guarantee the existence and uniqueness of solutions to the saddlepoint equation (4).

Lemma 22. *Under the assumptions in Theorem 1, the function $K_n''(s)$ is nonnegative on $(-\varepsilon, \varepsilon)$:*

$$K_n''(s) \geq 0 \quad \text{for all } s \in (-\varepsilon, \varepsilon) \text{ almost surely.} \tag{102}$$

Furthermore, it is uniformly bounded away from zero on a neighborhood of the origin, in the sense that for each $\delta > 0$, there exist $\eta > 0, s^* \in (0, \varepsilon/2)$ and $N \in \mathbb{N}_+$ such that

$$\mathbb{P}\left[\inf_{s \in [-s_*, s_*]} K_n''(s) \geq \eta\right] \geq 1 - \delta \quad \text{for all } n \geq N. \tag{103}$$

This lemma guarantees that the function $K_n'(s)$ is nondecreasing on $(-\varepsilon, \varepsilon)$ and increasing at a positive rate near the origin. To better illustrate the intuition, we refer the reader to Figure 7. Since $w_n \xrightarrow{\mathbb{P}} 0$, this implies that the saddlepoint equation (4) will have a solution for large enough n.

To be more precise, fix $\delta > 0$. By Lemma 22 and the fact that $w_n \xrightarrow{\mathbb{P}} 0$, let $\eta > 0$, $s_* \in (0, \varepsilon/2)$ and $N \in \mathbb{N}_+$ be such that

$$\mathbb{P}\left[\mathcal{E}_n\right] \geq 1 - \delta \quad \text{for all } n \geq N, \quad \text{where} \quad \mathcal{E}_n \equiv \left\{\inf_{s \in [-s_*, s_*]} K_n''(s) \geq \eta, |w_n| < s_* \eta\right\}. \tag{104}$$

On the event $\mathcal{E}_n \cap \mathcal{A}$, we can Taylor expand $K_n'(s)$ around $s = 0$ to obtain

$$K_n'(s) = K_n'(0) + s K_n''(\bar{s}) = s K_n''(\bar{s}) \quad \text{for} \quad |\bar{s}| \in (0, |s|), \tag{105}$$

where we have used the fact that

$$K_n'(0) = \frac{1}{n}\sum_{i=1}^n K_{in}'(0) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_{in} \mid \mathcal{F}_n] = 0. \tag{106}$$

It follows from the Taylor expansion (105) that, for $n \geq N$, we have

$$K_n'(-s_*) \leq -s_* \eta < w_n < s_* \eta \leq K_n'(s_*).$$

By the continuity of $K_n'(s)$ on the event $\mathcal{A}$ (Lemma 5), the intermediate value theorem implies that for each $n \geq N$, there exists a solution $\hat{s}_n \in (-s_*, s_*)$ to the saddlepoint equation (4) on the event $\mathcal{E}_n \cap \mathcal{A}$. Furthermore, for each $n \geq N$, this solution is unique

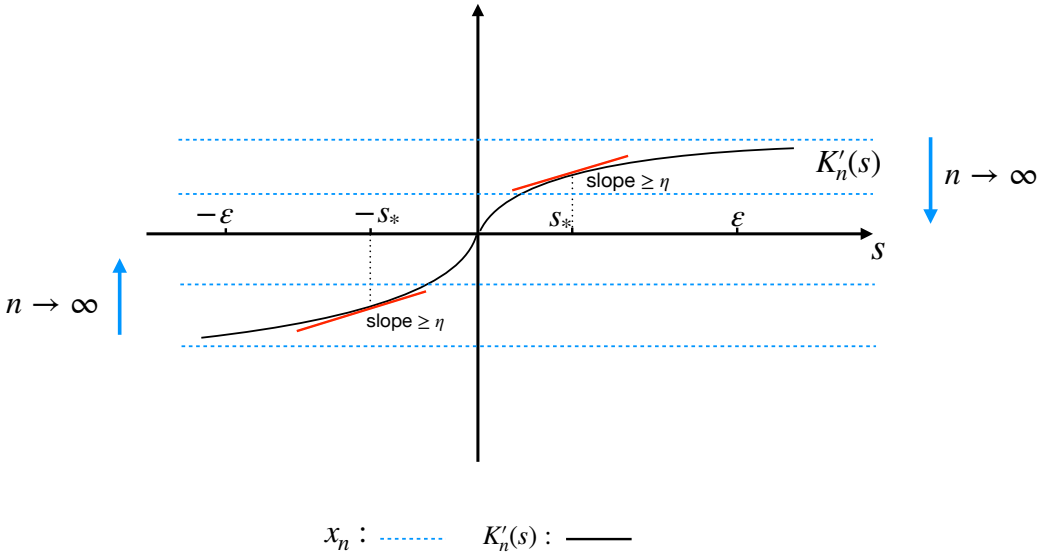

Figure 7: Illustration of the function $K_n'(s)$ for n large. The derivative of the function is nonnegative and strictly positive near the origin.

on $\mathcal{E}_n \cap \mathcal{A}$ because $K_n'(s)$ is strictly increasing on $[-s_*, s_*]$ and nondecreasing on the entire interval $[-\varepsilon/2, \varepsilon/2]$. Hence, we have shown that, for arbitrary $\delta > 0$, we have

$$\liminf_{n\to\infty} \mathbb{P}[|S_n| = 1] \geq \liminf_{n\to\infty} \mathbb{P}[\mathcal{E}_n \cap \mathcal{A}] \geq 1 - \delta. \tag{107}$$

Letting $\delta \to 0$ implies the first claim (41) of Theorem 1.

The following lemma records three properties of the saddlepoint $\hat{s}_n$, which will be useful in the remainder of the proof:

Lemma 23. *The saddlepoint $\hat{s}_n$ satisfies the following properties:*

$$\operatorname{sgn}(\hat{s}_n) = \operatorname{sgn}(w_n) \text{ almost surely}; \tag{108}$$

$$\hat{s}_n \xrightarrow{\mathbb{P}} 0; \tag{109}$$

$$K_n''(\hat{s}_n) = \Omega_{\mathbb{P}}(1). \tag{110}$$

G.2 Decomposing based on the sign of w_n

Since w_n is random, it can have uncertainty on the sign. This is a technical challenge since the uncertain sign of w_n will also make the signs of λ_n, r_n uncertain. We observe that the desired result (8) is implied by the following three statements, which

decompose the problem based on the sign of w_n:

$$\mathbb{1}(w_n > 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]}{1-\Phi(r_n)+\phi(r_n)\left\{\frac{1}{\lambda_n}-\frac{1}{r_n}\right\}}-1\right) \xrightarrow{\mathbb{P}} 0; \tag{111}$$

$$\mathbb{1}(w_n < 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]}{1-\Phi(r_n)+\phi(r_n)\left\{\frac{1}{\lambda_n}-\frac{1}{r_n}\right\}}-1\right) \xrightarrow{\mathbb{P}} 0; \tag{112}$$

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq 0 \mid \mathcal{F}_n\right] \xrightarrow{\mathbb{P}} \frac{1}{2}. \tag{113}$$

In the next two subsections, we verify statements (112) and (113), respectively. This will leave just the statement (111).

G.2.1 Verifying statement (112)

Before verifying statement (112), we state a lemma on the properties of the quantities λ_n and r_n that are necessary for proving the statement:

Lemma 24. *Under the assumptions of Theorem 1, r_n and λ_n are almost surely finite:*

$$|r_n| < \infty, |\lambda_n| < \infty, \ \ a.s. \tag{Finite}$$

Furthermore, the signs of w_n, r_n, and λ_n have the following relationships:

$$w_n > 0 \Rightarrow \lambda_n \geq 0, r_n \geq 0, \ \ a.s.; \tag{Sign1}$$

$$\mathbb{P}[w_n > 0 \ \textit{and} \ \lambda_n r_n = 0] \to 0; \tag{Sign2}$$

$$\mathbb{P}[\hat{s}_n \neq 0 \ \textit{and} \ \lambda_n r_n = 0] \to 0; \tag{Sign3}$$

$$r_n < 0 \Rightarrow \lambda_n \leq 0, \ \ \lambda_n < 0 \Rightarrow r_n \leq 0, \ \ a.s. \tag{Sign4}$$

Finally, r_n and λ_n satisfy the following convergence statements:

$$\frac{1}{\lambda_n} - \frac{1}{r_n} = o_{\mathbb{P}}(1); \tag{Rate1}$$

$$\frac{\lambda_n}{r_n} - 1 = o_{\mathbb{P}}(1); \tag{Rate2}$$

$$\mathbb{1}(r_n > 0, \lambda_n > 0)\frac{1}{r_n}\left(\frac{\lambda_n}{r_n} - 1\right) = o_{\mathbb{P}}(1); \tag{Rate3}$$

$$\mathbb{1}(\lambda_n \neq 0)\frac{1}{\lambda_n}\left(\frac{r_n}{\lambda_n} - 1\right) = o_{\mathbb{P}}(1); \tag{Rate4}$$

$$\frac{r_n}{\sqrt{n}} = o_{\mathbb{P}}(1). \tag{Rate5}$$

Now we claim that if the statement (111) holds, then we can derive the statement (112) by symmetry:

Lemma 25. *Suppose the assumptions of Theorem 1 hold and imply statement (111). We can apply the theorem to the triangular array* $\widetilde{W}_{in} \equiv -W_{in}$ *and set of cutoffs* $\widetilde{w}_n \equiv -w_n$ *to obtain that*

$$\mathbb{1}(\widetilde{w}_n > 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{W}_{in} \geq \widetilde{w}_n \mid \mathcal{F}_n\right]}{1-\Phi(\widetilde{r}_n)+\phi(\widetilde{r}_n)\left\{\frac{1}{\widetilde{\lambda}_n}-\frac{1}{\widetilde{r}_n}\right\}}-1\right) \xrightarrow{\mathbb{P}} 0,$$

where $\widetilde{r}_n = -r_n$ *and* $\widetilde{\lambda}_n = -\lambda_n$*. Then under conditions (Finite), (Sign1), (Sign4) and (Rate1), the following convergence statement holds:*

$$\mathbb{1}(w_n < 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} \geq w_n \mid \mathcal{F}_n\right]}{1-\Phi(r_n)+\phi(r_n)\left\{\frac{1}{\lambda_n}-\frac{1}{r_n}\right\}}-1\right) \xrightarrow{\mathbb{P}} 0. \tag{114}$$

G.2.2 Verifying statement (113)

To prove the statement (113), we first state a conditional central limit theorem:

Lemma 26 (Niu et al., 2024)**.** *Consider a sequence of* σ*-algebras* $\mathcal{F}_n$ *and probability measures* $\mathbb{P}_n$*. Let* W_{in} *be a triangular array of random variables, such that for each* n*,* W_{in} *are independent conditionally on* $\mathcal{F}_n$ *under* $\mathbb{P}_n$*. Let*

$$S_n^2 \equiv \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{\mathbb{P}_n}[W_{in} \mid \mathcal{F}_n].$$

If $\mathrm{Var}_{\mathbb{P}_n}[W_{in} \mid \mathcal{F}_n] < \infty$ *almost surely for each* i *and* n*, and for some* $\delta > 0$ *we have*

$$n^{-\delta/2}\frac{1}{S_n^{2+\delta}}\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{\mathbb{P}_n}[|W_{in} - \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n]|^{2+\delta} \mid \mathcal{F}_n] \xrightarrow{\mathbb{P}_n} 0, \tag{115}$$

then

$$\sup_{z\in\mathbb{R}}\left|\mathbb{P}_n\left[\sqrt{\frac{n}{S_n^2}}\frac{1}{n}\sum_{i=1}^n (W_{in} - \mathbb{E}_{\mathbb{P}_n}[W_{in} \mid \mathcal{F}_n]) \leq z \mid \mathcal{F}_n\right] - \Phi(z)\right| \xrightarrow{\mathbb{P}_n} 0.$$

We can apply this result to the variables W_{in} to get the following convergence statements:

Lemma 27. *Suppose the conditions in Theorem 1 hold. Then we have*

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left[\sqrt{\frac{n}{K_n''(0)}}\frac{1}{n}\sum_{i=1}^n W_{in} \leq t|\mathcal{F}_n\right] - \Phi(t)\right| \xrightarrow{\mathbb{P}} 0. \tag{116}$$

Moreover, for any sequence $y_n \in \mathcal{F}_n$*, we know*

$$\mathbb{P}\left[\sqrt{\frac{n}{K_n''(0)}}\frac{1}{n}\sum_{i=1}^n W_{in} = y_n|\mathcal{F}_n\right] \xrightarrow{\mathbb{P}} 0. \tag{117}$$

Setting $t = 0$ and $y_n = 0$ in Lemma 27, we obtain

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq 0 | \mathcal{F}_n\right] = 1 - \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \leq 0 | \mathcal{F}_n\right] + \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} = 0 | \mathcal{F}_n\right]$$
$$\xrightarrow{\mathbb{P}} 1 - \frac{1}{2} + 0 = \frac{1}{2},$$

which verifies (113).

G.3 Conditional Berry-Esseen bound on tilted summands

It remains to prove the statement (111) regarding the conditional tail probability $\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]$. This tail probability can be approximated using the conditional central limit theorem (Lemma 26). However, the central limit theorem is insufficiently accurate in the tails of the distribution. To overcome this challenge, we apply a normal approximation after exponential tilting, as is common in saddlepoint approximations (Robinson, 1982; Reid, 1988). The idea is to consider a probability distribution $\mathbb{P}_n$ over the space such that

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\mathbb{P}_n}[W_{in} \mid \mathcal{F}_n] = w_n. \tag{118}$$

Under such $\mathbb{P}_n$, the distribution of $\frac{1}{n}\sum_{i=1}^{n} W_{in}$ can be approximated as a normal with conditional mean w_n, allowing us to avoid approximating extreme tail probabilities. We can then undo the exponential tilting to approximate the desired tail probability under the original measure $\mathbb{P}$.

G.3.1 Exponential tilting

Given tilting parameter s, define a new probability measure $\mathbb{P}_{n,s}$ on the measurable space $(\Omega, \mathcal{F})$ via

$$\frac{\mathrm{d}\mathbb{P}_{n,s}}{\mathrm{d}\mathbb{P}} \equiv \prod_{i=1}^{n} \frac{\exp(sW_{in})}{\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n]}. \tag{119}$$

We employ a variant of tilting measure (119) based on a random tilting parameter $s_n \in \mathcal{F}_n$ that satisfies the criterion $\mathbb{P}[s_n \in (-\varepsilon, \varepsilon)] = 1$. The following lemma presents some properties of the tilted measure $\mathbb{P}_{n,s_n}$:

Lemma 28. *First, events in $\mathcal{F}_n$ are preserved under $\mathbb{P}_{n,s_n}$:*

$$\mathbb{P}_{n,s_n}[A_n] = \mathbb{P}[A_n] \quad \textit{for all} \quad A_n \in \mathcal{F}_n. \tag{120}$$

It follows that any random variable measurable with respect to $\mathcal{F}_n$ has the same distribution under $\mathbb{P}_{n,s_n}$ as under $\mathbb{P}$. Second, the random variables $\{W_{in}\}_{1\leq i\leq n}$ are independent conditionally on $\mathcal{F}_n$ under $\mathbb{P}_{n,s_n}$. Third, on the event $\mathcal{A}$, the conditional mean and variance of W_{in} under $\mathbb{P}_{n,s_n}$ are given by the first two derivatives of the conditional cumulant generating function K_{in}:

$$\mathbb{E}_{n,s_n}[W_{in} \mid \mathcal{F}_n] = K'_{in}(s_n) \textit{ and } \mathrm{Var}_{n,s_n}[W_{in} \mid \mathcal{F}_n] = K''_{in}(s_n) \textit{ for all } i \leq n, n \geq 1 \tag{121}$$

almost surely.

It follows from equation (121) that, almost surely,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{n,s_n}[W_{in} \mid \mathcal{F}_n] = K_n'(s_n) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{n,s_n}[W_{in} \mid \mathcal{F}_n] = K_n''(s_n).$$

To ensure the property (118), it suffices to take $\mathbb{P}_n \equiv \mathbb{P}_{n,\hat{s}_n}$, where $\hat{s}_n$ is the solution to the saddlepoint equation (4). Therefore, our next step is to construct a normal approximation for the average $\frac{1}{n}\sum_{i=1}^{n} W_{in}$ under the sequence of tilted probability measures $\mathbb{P}_{n,\hat{s}_n}$.

G.3.2 Conditional Berry-Esseen

It turns out that rate of the normal approximation is important to obtain a relative error bound, so we use the conditional Berry-Esseen theorem rather than the central limit theorem on the tilted summands.

Lemma 29 (Conditional Berry-Esseen theorem)**.** *Suppose $W_{1n}, \ldots, W_{nn}$ are independent random variables conditional on $\mathcal{F}_n$, under $\mathbb{P}_n$. If*

$$S_n^2 \equiv \frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}_{\mathbb{P}_n}[W_{in} \mid \mathcal{F}_n] = \Omega_{\mathbb{P}_n}(1) \tag{122}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\mathbb{P}_n}[|W_{in} - \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n]|^3|\mathcal{F}_n] = O_{\mathbb{P}_n}(1), \tag{123}$$

then

$$\sqrt{n}\sup_{t\in\mathbb{R}}\left|\mathbb{P}_n\left[\sqrt{\frac{n}{S_n^2}}\frac{1}{n}\sum_{i=1}^{n}(W_{in} - \mathbb{E}_{\mathbb{P}_n}[W_{in}|\mathcal{F}_n]) \le t|\mathcal{F}_n\right] - \Phi(t)\right| = O_{\mathbb{P}_n}(1).$$

Now, we wish to apply the Lemma 29 to the triangular array $\{W_{in}\}_{1\le i\le n, n\ge 1}$ under the sequence of tilted probability measures $\mathbb{P}_{n,\hat{s}_n}$. The following lemma shows that the requisite conditions are satisfied.

Lemma 30. *Under the assumptions of Theorem 1, the conditions (122) and (123) are satisfied by the sequence of probability measures $\mathbb{P}_n \equiv \mathbb{P}_{n,\hat{s}_n}$.*

Noting from equation (121) that $S_n^2 = K_n''(\hat{s}_n)$, we conclude from the conditional Berry-Esseen theorem that

$$\begin{aligned}
&\sqrt{n}\sup_{t\in\mathbb{R}}\left|\mathbb{P}_n\left[\sqrt{\frac{n}{K_n''(\hat{s}_n)}}\left(\frac{1}{n}\sum_{i=1}^{n}W_{in} - K_n'(\hat{s}_n)\right) \le t|\mathcal{F}_n\right] - \Phi(t)\right| \\
&\equiv \sqrt{n}\sup_{t\in\mathbb{R}}\left|\mathbb{P}_n\left[\widetilde{Z}_n \le t \mid \mathcal{F}_n\right] - \Phi(t)\right| \\
&= O_{\mathbb{P}_n}(1),
\end{aligned} \tag{124}$$

where we have denoted by $\widetilde{Z}_n$ the quantity converging to the standard normal distribution. Note that $\widetilde{Z}_n$ is not exactly the same as

$$Z_n \equiv \sqrt{\frac{n}{K_n''(\hat{s}_n)}} \left(\frac{1}{n} \sum_{i=1}^{n} W_{in} - w_n \right), \tag{125}$$

since it is possible that $K_n'(\hat{s}_n) \neq w_n$. Since the probability of this event is tending to zero (41), we find that

$$\begin{aligned}
&\sqrt{n} \sup_{t \in \mathbb{R}} |\mathbb{P}_n[Z_n \leq t | \mathcal{F}_n] - \Phi(t)| \\
&= \mathbb{1}(K_n'(\hat{s}_n) = w_n) \sqrt{n} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_n[\widetilde{Z}_n \leq t | \mathcal{F}_n] - \Phi(t) \right| \\
&\quad + \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) \sqrt{n} \sup_{t \in \mathbb{R}} |\mathbb{P}_n[Z_n \leq t | \mathcal{F}_n] - \Phi(t)| \\
&\leq \sqrt{n} \sup_{t \in \mathbb{R}} \left| \mathbb{P}_n[\widetilde{Z}_n \leq t | \mathcal{F}_n] - \Phi(t) \right| + \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) \sqrt{n} \\
&= O_{\mathbb{P}_n}(1) + o_{\mathbb{P}_n}(1) = O_{\mathbb{P}_n}(1).
\end{aligned}$$

By conclusion (120) from Lemma 28 and the measurability with respect to $\mathcal{F}_n$ of the quantity $\sqrt{n} \sup_{t \in \mathbb{R}} |\mathbb{P}_n[Z_n \leq t | \mathcal{F}_n] - \Phi(t)|$, it follows that

$$\sqrt{n} \sup_{t \in \mathbb{R}} |\mathbb{P}_n[Z_n \leq t | \mathcal{F}_n] - \Phi(t)| = O_{\mathbb{P}}(1). \tag{126}$$

Therefore, we have provided a normal approximation for the average $\frac{1}{n} \sum_{i=1}^{n} W_{in}$ under the sequence of tilted probability measures $\mathbb{P}_{n,\hat{s}_n}$. Next, we undo the exponential tilting to approximate the desired tail probability under the original measure $\mathbb{P}$.

G.4 Gaussian integral approximation after tilting back

G.4.1 Tilting back to the original measure

The following lemma helps connect the tilted measure to the original one, allowing us to interchange the order of the tilting and the conditioning:

Lemma 31.

$$\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} W_{in} \geq w_n \;\middle|\; \mathcal{F}_n \right] = \mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}\left(\frac{1}{n} \sum_{i=1}^{n} W_{in} \geq w_n \right) \frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}} \mid \mathcal{F}_n \right], \tag{127}$$

where $\hat{s}_n$ is the solution to the saddlepoint equation (40) for each n.

To evaluate the right-hand side of equation (127), we first note that

$$\begin{aligned}\frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}} &= \prod_{i=1}^{n} \frac{\mathbb{E}[\exp(\hat{s}_n W_{in})|\mathcal{F}_n]}{\exp(\hat{s}_n W_{in})} \\ &= \exp\left(n\left(K_n(\hat{s}_n) - \hat{s}_n \frac{1}{n}\sum_{i=1}^{n} W_{in}\right)\right) \\ &= \exp\left(n(K_n(\hat{s}_n) - \hat{s}_n w_n) - \hat{s}_n\sqrt{nK_n''(\hat{s}_n)}\sqrt{\frac{n}{K_n''(\hat{s}_n)}}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} - w_n\right)\right) \\ &\equiv \exp\left(-\frac{1}{2}r_n^2 - \lambda_n Z_n\right),\end{aligned}$$

recalling λ_n and r_n from equation (5) and Z_n (125) the quantity converging to normality (126). This allows us to rewrite the probability of interest as

$$\begin{aligned}\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right] &= \mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}(Z_n \geq 0)\exp\left(-\frac{1}{2}r_n^2 - \lambda_n Z_n\right) \mid \mathcal{F}_n\right] \\ &= \exp\left(-\frac{1}{2}r_n^2\right)\mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}(Z_n \geq 0)\exp\left(-\lambda_n Z_n\right) \mid \mathcal{F}_n\right].\end{aligned} \tag{128}$$

Therefore, we have

$$\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n \mid \mathcal{F}_n\right]}{1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}} = \frac{\mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}(Z_n \geq 0)\exp\left(-\lambda_n Z_n\right) \mid \mathcal{F}_n\right]}{\exp\left(\frac{1}{2}r_n^2\right)\left(1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}\right)} \equiv \frac{D_n}{U_n}. \tag{129}$$

Hence, we have simplified the desired statement (111) to

$$\mathbb{1}(w_n > 0)\left(\frac{D_n}{U_n} - 1\right) \xrightarrow{\mathbb{P}} 0. \tag{130}$$

G.4.2 Reduction to a Gaussian integral approximation

Next, we exploit the convergence of Z_n to normality (126) to replace the numerator D_n with a Gaussian integral:

Lemma 32. *For sequences Z_n and λ_n of random variables, we have*

$$\begin{aligned}&\mathbb{1}(\lambda_n \geq 0)\left|\mathbb{E}_{\mathbb{P}_n}\left[\mathbb{1}(Z_n \geq 0)\exp\left(-\lambda_n Z_n\right) \mid \mathcal{F}_n\right] - \int_0^\infty \exp(-\lambda_n z)\phi(z)dz\right| \\ &\quad \leq 2\mathbb{1}(\lambda_n \geq 0)\sup_{t\in\mathbb{R}}\left|\mathbb{P}_n\left[Z_n \leq t|\mathcal{F}_n\right] - \Phi(t)\right|.\end{aligned} \tag{131}$$

almost surely.

We would like to combine the result of this lemma with the convergence of Z_n to normality (126) to reduce the desired statement (130) to a Gaussian integral approximation. Before doing so, we first state a result that we will use to show that the difference between D_n and the Gaussian integral $\int_0^\infty \exp(-\lambda_n z)\phi(z)dz$ is negligible, even after dividing by U_n:

Lemma 33. *Under conditions* (Finite), (Sign4) *and* (Rate1), *we have*

$$r_n \geq 0 \Rightarrow U_n \neq 0 \text{ almost surely; } \mathbb{P}[r_n < 0 \text{ and } U_n = 0] \to 0. \tag{132}$$

Under conditions (Rate1), (Rate2) *and* (Rate5), *we have*

$$\frac{\mathbb{1}(r_n \geq 0)}{\sqrt{n}U_n} = o_{\mathbb{P}}(1). \tag{133}$$

Therefore, we have

$$\begin{aligned}
&\mathbb{1}(w_n > 0) \left| \frac{D_n}{U_n} - \frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} \right| \\
&\leq \mathbb{1}(r_n \geq 0, \lambda_n \geq 0) \left| \frac{D_n}{U_n} - \frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} \right| && \text{by (Sign1)} \\
&\leq \frac{2\mathbb{1}(r_n \geq 0, \lambda_n \geq 0) \sup_{t\in\mathbb{R}} |\mathbb{P}_n [Z_n \leq t \mid \mathcal{F}_n] - \Phi(t)|}{|U_n|} && \text{by Lemma 32} \\
&\leq \frac{\mathbb{1}(r_n \geq 0)}{\sqrt{n}|U_n|} O_{\mathbb{P}}(1) && \text{by (126)} \\
&= o_{\mathbb{P}}(1) O_{\mathbb{P}}(1) && \text{by (133)} \\
&= o_{\mathbb{P}}(1),
\end{aligned}$$

Therefore, it suffices to show that

$$\mathbb{1}(w_n > 0) \left(\frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} - 1 \right) = o_{\mathbb{P}}(1). \tag{134}$$

By statements (Sign1), (Sign2), and (132), it suffices to show the Gaussian integral approximation

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \left(\frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} - 1 \right) = o_{\mathbb{P}}(1), \tag{135}$$

which is stated in the following lemma:

Lemma 34. *Under conditions* (Finite), (Rate1), (Rate2), *and* (Rate3), *the Gaussian integral approximation* (135) *holds.*

This completes the proof Theorem 1.

H Proof of Proposition 2

From (8), it suffices to prove

$$\frac{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}}{\exp\left(\frac{\lambda_n^2 - r_n^2}{2} \right) (1 - \Phi(\lambda_n))} = 1 + o_{\mathbb{P}}(1).$$

On the event $\hat{s}_n = 0$, we know the claim is correct. Therefore, we only need to consider the event $\hat{s}_n \neq 0$. Equivalently, it suffices to show

$$\mathbb{1}(\hat{s}_n \neq 0) \left(\frac{\exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n))}{\exp\left(\frac{\lambda_n^2}{2}\right)(1-\Phi(\lambda_n))} - 1 \right) \equiv \mathbb{1}(\hat{s}_n \neq 0) \left(\frac{h(r_n)}{h(\lambda_n)} - 1 \right) = o_{\mathbb{P}}(1) \tag{136}$$

and

$$\mathbb{1}(\hat{s}_n \neq 0) \frac{\frac{1}{\lambda_n} - \frac{1}{r_n}}{\exp(\lambda_n^2/2)(1-\Phi(\lambda_n))} = \mathbb{1}(\hat{s}_n \neq 0) \frac{1 - \frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1). \tag{137}$$

We prove the statements (136)-(137) subsequently.

H.1 Proof of statement (136)

Since $h(x) = \exp(x^2/2)(1-\Phi(x))$ is smooth, then by Taylor's expansion, we have

$$\frac{h(r_n)}{h(\lambda_n)} = \frac{h(\lambda_n) + h'(\tilde{r}_n)(r_n - \lambda_n)}{h(\lambda_n)} = 1 + \frac{h'(\tilde{r}_n)(r_n - \lambda_n)}{h(\lambda_n)}$$

where $\tilde{r}_n$ is the point between r_n and λ_n. Now we intestigate $h'(x)$. We compute

$$h'(x) = x \exp(x^2/2)(1-\Phi(x)) - \frac{1}{\sqrt{2\pi}}.$$

By Lemma 17, we know

$$|h'(x)| \leq \frac{2}{\sqrt{2\pi}} \frac{1}{x^2} \leq \frac{1}{x^2}.$$

Then since both event $r_n < 0, \lambda_n > 0$ and event $r_n > 0, \lambda_n < 0$ happen with probability zero, we have $1/\tilde{r}_n^2 \in [\min\{1/r_n^2, 1/\lambda_n^2\}, \max\{1/r_n^2, 1/\lambda_n^2\}]$. Therefore, we have

$$\left| \frac{h'(\tilde{r}_n)(r_n - \lambda_n)}{h(\lambda_n)} \right| \leq \frac{1}{\tilde{r}_n^2} \left| \frac{r_n - \lambda_n}{h(\lambda_n)} \right| \leq \left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2} \right) \left| \frac{r_n - \lambda_n}{h(\lambda_n)} \right| = \left(1 + \frac{\lambda_n^2}{r_n^2} \right) \left| \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} \right|.$$

Thus in order to prove (136), it suffices to show, by the sign condition (Sign3),

$$\mathbb{1}(\lambda_n \neq 0) \left(1 + \frac{\lambda_n^2}{r_n^2} \right) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

The following lemma shows that the above statement is correct:

Lemma 35. *Under conditions (Rate2) and (Rate4), we have*

$$\mathbb{1}(\lambda_n \neq 0) \left(1 + \frac{\lambda_n^2}{r_n^2} \right) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

H.2 Proof of statement (137)

By the sign condition (Sign3), it suffices to prove

$$\mathbb{1}(\lambda_n \neq 0)\frac{1-\frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

The following lemma shows that the above statement is correct:

Lemma 36. *Under conditions (Rate1) and (Rate2), we have*

$$\mathbb{1}(\lambda_n \neq 0)\frac{1-\frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

I Proofs of supporting lemmas for Theorem 1

In this section, we first state two lemmas that reduce the condition of Theorem 1 to several conditions on the CGF. Then we prove the supporting lemmas for Theorem 1 based on the reduced conditions.

Lemma 37. *Suppose Assumption 1 or Assumption 2 holds. Then, the following statements hold:*

$$\sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}(K_{in}''(s))^2 = O_{\mathbb{P}}(1); \tag{138}$$

$$\frac{1}{n}\sum_{i=1}^{n} K_{in}'''(0) = O_{\mathbb{P}}(1); \tag{139}$$

$$\sup_{s\in(-\varepsilon,\varepsilon)} \left|\frac{1}{n}\sum_{i=1}^{n} K_{in}''''(s)\right| = O_{\mathbb{P}}(1). \tag{140}$$

Lemma 38. *Suppose Assumption 1 or Assumption 2 holds. Then condition (7) implies*

$$\frac{1}{n}\sum_{i=1}^{n} K_{in}''(0) = \Omega_{\mathbb{P}}(1). \tag{141}$$

I.1 Proof of Lemma 5

We prove claim (38) and (39) separately.

Proof of claim (38): We consider two cases: CSE distribution and CCS distribution.

Case 1: CSE distribution By Lemma 21, we know

$$\mathbb{P}\left[K_{in}(s) \leq \lambda_n s^2,\ \forall s \in \left(-\frac{1}{\gamma}, \frac{1}{\gamma}\right)\right] = 1$$

where

$$\lambda_n \equiv \frac{\sqrt{6!4^6}(1+\theta_n)}{24\beta^2} + \frac{16(1+\theta_n)}{\beta^2},\ \gamma = \frac{4}{\beta}.$$

Since $\theta_n < \infty$ almost surely, we know condition (38) holds with $\varepsilon = 1/(2\gamma) = \beta/8$.

Case 2: CCS distribution By the definition of CCS distribution and the definition of regular conditional distribution, we have for almost every $\omega \in \Omega$,

$$\mathbb{P}\left[\mathrm{Supp}(\kappa_{in}(\omega,\cdot)) \in [-\nu_{in}(\omega), \nu_{in}(\omega)]\right] = 1.$$

Then we have for almost every $\omega \in \Omega$,

$$\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n](\omega) = \int \exp(sx)\mathrm{d}\kappa_{in}(\omega, x) \leq \exp(\nu_{in}(\omega)) < \infty, \ \forall s \in (-1, 1)$$

where the last inequality is due to the assumption $\nu_{in} < \infty$ almost surely. Therefore, condition (38) holds with $\varepsilon = 1$.

Proof of claim (39): By Lemma 19, it suffices to prove the following lemma.

Lemma 39. *On the event $\mathcal{A}$,*

$$\mathbb{E}[|W_{in}|^p \exp(sW_{in})|\mathcal{F}_n] < \infty, \ \forall s \in (-\varepsilon, \varepsilon), \ \forall i \in \{1, \ldots, n\}, n \geq 1 \text{ and } p \in \mathbb{N}.$$

Proof of Lemma 39 is postponed to Appendix I.18.

I.2 Proof of Lemma 22

The claim (102) holds because by Lemma 45, on the event $\mathcal{A}$ we have, for each $s \in (-\varepsilon, \varepsilon)$,

$$K_n''(s) = \frac{1}{n}\sum_{i=1}^n K_{in}''(s) = \frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{n,s}[W_{in} \mid \mathcal{F}_n] \geq 0.$$

Next, we verify claim (103). To this end, fix $\delta > 0$. By assumptions (141), (139), and (140), there exist $\eta, M > 0$ and $N \geq 1$ be such that for all $n \geq N$,

$$\mathbb{P}[K_n''(0) < 2\eta] < \delta/3, \quad \mathbb{P}[|K_n'''(0)| > M] < \delta/3, \quad \mathbb{P}\left[\sup_{s\in(-\varepsilon,\varepsilon)} |K_n''''(s)| > M\right] < \delta/3.$$

Define

$$s_* \equiv \min(\eta/(2M), \sqrt{\eta/M}, \varepsilon/2).$$

On the event $\mathcal{A}$, Lemma 5 guarantees that we can we Taylor expand $K_n''(s)$ around $s = 0$ to obtain

$$K_n''(s) = K_n''(0) + sK_n'''(0) + \frac{1}{2}s^2 K_n''''(\bar{s})$$

for some $|\bar{s}| \leq |s|$. Therefore, for all $n \geq N$, we have

$$\begin{aligned} 1 - \delta &< \mathbb{P}\left[\mathcal{A}, K_n''(0) \geq 2\eta, |K_n'''(0)| \leq M, \sup_{s\in(-\varepsilon,\varepsilon)} |K_n''''(s)| \leq M\right] \\ &\leq \mathbb{P}\left[\inf_{s\in[-s_*,s_*]} K_n''(s) \geq 2\eta - s_*M - \frac{1}{2}s_*^2 M \geq \eta\right], \end{aligned}$$

which verifies the claim (103) and completes the proof.

I.3 Proof of Lemma 23

Proof of (108): Suppose $|S_n| = 1$. Because K_n' is almost surely nondecreasing on $(-\varepsilon, \varepsilon)$ (102) and $K_n'(0) = 0$ (106), the identity $K_n'(\hat{s}_n) = w_n$ implies that $\operatorname{sgn}(\hat{s}_n) = \operatorname{sgn}(w_n)$. When $|S_n| \neq 1$, by the definition of $\hat{s}_n$ (40), we have $\operatorname{sgn}(\hat{s}_n) = \operatorname{sgn}(w_n)$. This completes the proof.

Proof of (109): Fix $\gamma, \delta > 0$. By Lemma 22, there exist $\eta > 0$, $s_* \in (0, \varepsilon/2)$, and $N \in \mathbb{N}_+$ such that

$$\mathbb{P}\left[\inf_{s\in[-s_*,s_*]} K_n''(s) \geq \eta\right] \geq 1 - \delta/2 \quad \text{for all} \quad n \geq N.$$

By increasing N if necessary, the fact that $w_n \xrightarrow{\mathbb{P}} 0$ implies that

$$\mathbb{P}[|w_n| \leq \eta \min(\gamma, s_*)] \geq 1 - \delta/2 \quad \text{for all} \quad n \geq N.$$

Define the event

$$\mathcal{E}_n' \equiv \left\{\inf_{s\in[-s_*,s_*]} K_n''(s) \geq \eta, |w_n| \leq \eta \min(\gamma, s_*)\right\}$$

On the event $\mathcal{E}_n' \cap \mathcal{A}$, the Taylor expansion (105) gives

$$|K_n'(s)| \geq |s|\eta \quad \text{for all} \quad s \in [-s_*, s_*] \text{ and all } n \geq N.$$

Hence, $|w_n| \leq \eta s_* \leq \min(-K_n'(-s_*), K_n'(s_*))$, implying $w_n \in [K_n'(-s_*), K_n'(s_*)]$, so the saddlepoint equation has a solution $\hat{s}_n$ such that $K_n'(\hat{s}_n) = w_n$ and $|\hat{s}_n| \leq s_*$. Therefore, on the event $\mathcal{E}_n' \cap \mathcal{A}$, we have

$$|\hat{s}_n|\eta \leq |K_n'(\hat{s}_n)| = |w_n| \leq \eta\gamma \quad \Longrightarrow \quad |\hat{s}_n| \leq \gamma.$$

It follows that

$$\mathbb{P}[|\hat{s}_n| \leq \gamma] \geq \mathbb{P}[\mathcal{E}_n' \cap \mathcal{A}] \geq 1 - \delta \quad \text{for all} \quad n \geq N,$$

which shows that $\hat{s}_n \xrightarrow{\mathbb{P}} 0$, as desired.

Proof of (110): By the argument following the statement of Lemma 22, for any δ there is an $\eta > 0$ and $N \in \mathbb{N}_+$ such that $\mathbb{P}[K_n''(\hat{s}_n) \geq \eta] \geq 1 - \delta$ for all $n \geq N$. This shows that $K_n''(\hat{s}_n) = \Omega_{\mathbb{P}}(1)$, as desired.

I.4 Proof of Lemma 24

Proof of Lemma 24. We prove the claims separately.

Verification of (Finite). Since $w_n \in (-\infty, \infty)$, together with Lemma 5 guaranteeing that $K_n(s), K_n'(s), K_n''(s) \in (-\infty, \infty), \forall s \in (-\varepsilon, \varepsilon)$ almost surely and definition of $\hat{s}_n$ such that $|\hat{s}_n| < \varepsilon$, we have

$$\lambda_n^2 = |n\hat{s}_n K_n''(\hat{s}_n)| < n\varepsilon |K_n''(\hat{s}_n)| < \infty, \; r_n^2 \leq \max\{1, |2n(\hat{s}_n w_n - K_n(\hat{s}_n))|\} < \infty.$$

Verification of (Sign1)**.** This claim follows from conclusion (108) of Lemma 23 and the definitions of r_n and λ_n in equation (5).

Verification of (Sign4)**.** This is true by definition of r_n and λ_n. This completes the proof.

Verification of (Sign2)**,** (Sign3)**,** (Rate1)**,** (Rate2)**,** (Rate3)**,** (Rate4) **and** (Rate5)**:** We present a useful lemma.

Lemma 40 (Asymptotic estimate of λ_n and r_n)**.** *Under the assumptions of Theorem 1, the followings are true*

$$\frac{r_n^2}{n} = o_{\mathbb{P}}(1); \tag{142}$$

$$\frac{\lambda_n}{r_n} = 1 + \hat{s}_n O_{\mathbb{P}}(1); \tag{143}$$

$$\frac{r_n}{\lambda_n} = 1 + \hat{s}_n O_{\mathbb{P}}(1); \tag{144}$$

$$\frac{1}{\lambda_n} - \frac{1}{r_n} = o_{\mathbb{P}}(1); \tag{145}$$

$$\mathbb{1}(r_n > 0, \lambda_n > 0)\frac{1}{r_n}\left(\frac{\lambda_n}{r_n} - 1\right) = o_{\mathbb{P}}(1); \tag{146}$$

$$\mathbb{1}(\lambda_n \neq 0)\frac{1}{\lambda_n}\left(\frac{r_n}{\lambda_n} - 1\right) = o_{\mathbb{P}}(1); \tag{147}$$

$$\mathbb{P}[w_n > 0 \text{ and } \lambda_n r_n \leq 0] \to 0. \tag{148}$$

$$\mathbb{P}[\hat{s}_n \neq 0 \text{ and } \lambda_n r_n \leq 0] \to 0. \tag{149}$$

Verification of (Sign2)**:** (148) verifies (Sign2).

Verification of (Sign3)**:** (149) verifies (Sign3).

Verification of (Rate1)**:** (145) verifies (Rate1).

Verification of (Rate2)**:** Since $\hat{s}_n \xrightarrow{\mathbb{P}} 0$, we know (143) implies

$$\frac{\lambda_n}{r_n} = 1 + o_{\mathbb{P}}(1)$$

which verifies (Rate2).

Verification of (Rate3)**:** (146) verifies (Rate3).

Verification of (Rate4)**:** (147) verifies (Rate4).

Verification of (Rate5)**:** (142) verifies (Rate5).

□

I.5 Proof of Lemma 25

We can apply the theorem to the triangular array $\widetilde{W}_{in} \equiv -W_{in}$ and set of cutoffs $\widetilde{w}_n \equiv -w_n$, since the theorem assumptions are invariant to the signs of W_{in} and x_{in}. Therefore, we get the result

$$\mathbb{1}(\widetilde{w}_n > 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{W}_{in} \geq \widetilde{w}_n \mid \mathcal{F}_n\right]}{1-\Phi(\widetilde{r}_n)+\phi(\widetilde{r}_n)\left\{\frac{1}{\widetilde{\lambda}_n}-\frac{1}{\widetilde{r}_n}\right\}}-1\right) \xrightarrow{\mathbb{P}} 0,$$

where we claim that $\widetilde{r}_n = -r_n$ and $\widetilde{\lambda}_n = -\lambda_n$. To see this, we define

$$\widetilde{K}_{in}(s) \equiv \log \mathbb{E}\left[\exp(s\widetilde{W}_{in})|\mathcal{F}_n\right],\ \widetilde{K}_n(s) \equiv \frac{1}{n}\sum_{i=1}^n \widetilde{K}_{in}(s) = \frac{1}{n}\sum_{i=1}^n K_{in}(-s) = K_n(-s).$$

Then, consider the saddlepoint equation for $\widetilde{w}_n$:

$$\widetilde{K}_n'(s) = \widetilde{x}_n. \tag{150}$$

Furthermore, we define

$$\widetilde{S}_n \equiv \{s \in [-\varepsilon/2, \varepsilon/2] : \widetilde{K}_n'(s) = \widetilde{x}_n\}.$$

Then we write the solution $\widetilde{s}_n$ to the saddlepoint equation (150) according to the definition of $\hat{s}_n$ as in (40)

$$\widetilde{s}_n = \begin{cases} \text{the single element of } \widetilde{S}_n & \text{if } |\widetilde{S}_n| = 1; \\ \frac{\varepsilon}{2}\text{sgn}(\widetilde{x}_n) & \text{otherwise.} \end{cases}$$

Then we argue that $\widetilde{s}_n = -\hat{s}_n$. This is because given $\hat{s}_n$ uniquely solves (4), we know $-\hat{s}_n$ uniquely solves (150). Similarly, whenever $\widetilde{s}_n$ uniquely solves (150), we know $-\widetilde{s}_n$ uniquely solves (4). Therefore, we have $\widetilde{s}_n = -\hat{s}_n$. Then recall the definition

$$\widetilde{\lambda}_n \equiv \sqrt{n}\widetilde{s}_n\widetilde{K}_n''(\widetilde{s}_n),\ \widetilde{r}_n \equiv \begin{cases} \text{sgn}(\widetilde{s}_n)\sqrt{2n(\widetilde{s}_n\widetilde{w}_n - \widetilde{K}_n(\widetilde{s}_n))} & \text{if } \widetilde{s}_n\widetilde{w}_n - \widetilde{K}_n(\widetilde{s}_n) \geq 0; \\ \text{sgn}(\widetilde{s}_n) & \text{otherwise,} \end{cases}.$$

Since $\widetilde{K}_n''(-s) = K_n''(s), \widetilde{K}_n(-s) = K_n(s)$ and $\widetilde{x}_n = -w_n$, we know $\widetilde{\lambda}_n = -\lambda_n$ and $\widetilde{r}_n = -r_n$. Therefore, we have

$$\begin{aligned}
&\mathbb{1}(\widetilde{w}_n > 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{W}_{in} \geq \widetilde{w}_n \mid \mathcal{F}_n\right]}{1-\Phi(\widetilde{r}_n)+\phi(\widetilde{r}_n)\left\{\frac{1}{\widetilde{\lambda}_n}-\frac{1}{\widetilde{r}_n}\right\}}-1\right) \\
&= \mathbb{1}(w_n < 0)\left(\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} \leq w_n \mid \mathcal{F}_n\right]}{1-\Phi(-r_n)+\phi(r_n)\left\{\frac{1}{r_n}-\frac{1}{\lambda_n}\right\}}-1\right) \\
&= \mathbb{1}(w_n < 0)\left(\frac{1-\Phi(r_n)+\phi(r_n)\left\{\frac{1}{\lambda_n}-\frac{1}{r_n}\right\}-\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} > w_n \mid \mathcal{F}_n\right]}{\Phi(r_n)+\phi(r_n)\left\{\frac{1}{r_n}-\frac{1}{\lambda_n}\right\}}\right) \xrightarrow{\mathbb{P}} 0.
\end{aligned} \tag{151}$$

Note the demoninator in (151) is not what we want and we would like to change it to $1-\Phi(r_n)+\phi(r_n)\{\frac{1}{\lambda_n}-\frac{1}{r_n}\}$. Now, we need the following lemma to proceed.

Lemma 41. *Suppose the assumptions of Theorem 1 hold. Then* (Finite), (Sign1), (Sign4), (Rate1) *conditions are true by Lemma 24. Furthermore, we have*

1.

$$\mathbb{1}(w_n < 0) \left| \frac{\Phi(r_n) + \phi(r_n) \left\{ \frac{1}{r_n} - \frac{1}{\lambda_n} \right\}}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} \right| \leq 1 + o_{\mathbb{P}}(1); \tag{152}$$

2.

$$\mathbb{1}(w_n < 0) \frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} = o_{\mathbb{P}}(1). \tag{153}$$

Now, guaranteed by (152) in Lemma 41, we multiply both sides of the last statement as in (151) by $\mathbb{1}(w_n < 0) \frac{\Phi(r_n) + \phi(r_n) \left\{ \frac{1}{r_n} - \frac{1}{\lambda_n} \right\}}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}}$ and rearrange to obtain that

$$\mathbb{1}(w_n < 0) \left(\frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} > w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} - 1 \right) \xrightarrow{\mathbb{P}} 0.$$

This is almost what we want (112), except the inequality in the numerator is strict. To address this, we note we have proved (153) in Lemma 41 that

$$\mathbb{1}(w_n < 0) \frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} \xrightarrow{\mathbb{P}} 0.$$

Putting together the preceding two displays, we conclude that

$$\begin{aligned}
&\mathbb{1}(w_n < 0) \left(\frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} \geq w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} - 1 \right) \\
&= \mathbb{1}(w_n < 0) \left(\frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} > w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} - 1 + \frac{\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n \right]}{1 - \Phi(r_n) + \phi(r_n) \left\{ \frac{1}{\lambda_n} - \frac{1}{r_n} \right\}} \right) \\
&\xrightarrow{\mathbb{P}} 0.
\end{aligned}$$

I.6 Proof of Lemma 27

Proof of Lemma 27. We first prove the first claim.

Proof of (116)**:** We apply Lemma 26 to prove the result. It suffices to show

1.

$$\mathrm{Var}[W_{in} | \mathcal{F}_n] < \infty;$$

2.

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|W_{in}-\mathbb{E}[W_{in}|\mathcal{F}_n]|^3|\mathcal{F}_n]=O_{\mathbb{P}}(1),\ \ K_n''(0)=\Omega_{\mathbb{P}}(1).$$

For the first claim, we know

$$\mathrm{Var}[W_{in}|\mathcal{F}_n]=K_{in}''(0)<\infty$$

almost surely by Lemma 5. For the second claim, we claim it suffices to prove

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|W_{in}-\mathbb{E}[W_{in}|\mathcal{F}_n]|^4|\mathcal{F}_n]=O_{\mathbb{P}}(1),\ \ K_n''(0)=\Omega_{\mathbb{P}}(1).$$

This is because, intuitively, we can upper bound the lower moment by the higher moment. Lemma 15 provides a formal result for such intuition. Applying Lemma 15 with $p=3, q=4$ we know the claim is true. By the expression of the fourth central moment in terms of the second and fourth cumulant, we have

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|W_{in}-\mathbb{E}[W_{in}|\mathcal{F}_n]|^4|\mathcal{F}_n]=\frac{1}{n}\sum_{i=1}^{n}\left\{K_{in}^{(4)}(0)+3(K_{in}''(0))^2\right\}=O_{\mathbb{P}}(1)$$

guaranteed by assumptions (138) and (140). $K_n''(0)=\Omega_{\mathbb{P}}(1)$ is guaranteed by assumption (141). Thus by Lemma 26, we know

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^{n}W_{in}\le t|\mathcal{F}_n\right]-\Phi(t)\right|\xrightarrow{\mathbb{P}}0.$$

Proof of (117): Fix $\delta>0$. Then we can bound

$$\begin{aligned}\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^{n}W_{in}=y_n|\mathcal{F}_n\right]&\le\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^{n}W_{in}\in(y_n-\delta,y_n+\delta]|\mathcal{F}_n\right]\\&\equiv P((y_n-\delta,y_n+\delta])\end{aligned}$$

where

$$P(A)\equiv\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^{n}W_{in}\in A|\mathcal{F}_n\right],\ A\subset\mathbb{R}.$$

Furthermore we have

$$\begin{aligned}P((y_n-\delta,y_n+\delta])\le&|P((-\infty,y_n+\delta])-\Phi(y_n+\delta)|+|P((-\infty,y_n-\delta])-\Phi(y_n-\delta)|\\&+|\Phi(y_n+\delta)-\Phi(y_n-\delta)|.\end{aligned}$$

By (116) and the Lipschitz continuity of $\Phi(x)$, we can bound

$$\begin{aligned}P((y_n-\delta,y_n+\delta])&\le 2\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^{n}W_{in}\le t|\mathcal{F}_n\right]-\Phi(t)\right|+\sup_{x\in\mathbb{R}}\phi(x)2\delta\\&=\sup_{x\in\mathbb{R}}\phi(x)2\delta+o_{\mathbb{P}}(1).\end{aligned}$$

Since $\sup_{x\in\mathbb{R}} \phi(x) \le 1/\sqrt{2\pi}$, we know

$$\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^n W_{in} = y_n|\mathcal{F}_n\right] \le P((y_n-\delta, y_n+\delta]) = o_{\mathbb{P}}(1) + \frac{2\delta}{\sqrt{2\pi}}.$$

We can take δ arbitrarily small so that we obtain

$$\mathbb{P}\left[\frac{1}{\sqrt{nK_n''(0)}}\sum_{i=1}^n W_{in} = y_n|\mathcal{F}_n\right] = o_{\mathbb{P}}(1).$$

□

I.7 Proof of Lemma 28

Proof of Lemma 28. To prove the statement (120), note that

$$\begin{aligned}
\mathbb{P}_{n,s_n}[A_n] &= \mathbb{E}\left[\mathbb{1}(A_n)\prod_{i=1}^n \frac{\exp(s_nW_{in})}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mathbb{1}(A_n)\prod_{i=1}^n \frac{\exp(s_nW_{in})}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\mid\mathcal{F}_n\right]\right] \\
&= \mathbb{E}\left[\mathbb{1}(A_n)\prod_{i=1}^n \frac{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\right] = \mathbb{P}[A_n].
\end{aligned}$$

Next, we compute for each $A_n \in \mathcal{F}_n$ and $B_1,\dots,B_n \subseteq \mathcal{B}(\mathbb{R})$ that

$$\begin{aligned}
&\mathbb{P}_{n,s_n}[W_{1n}\in B_1,\dots,W_{nn}\in B_n, A_n] \\
&= \mathbb{E}\left[\mathbb{1}(W_{1n}\in B_1,\dots,W_{nn}\in B_n, A_n)\prod_{i=1}^n \frac{\exp(s_nW_{in})}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\right] \\
&= \mathbb{E}\left[\mathbb{1}(A_n)\prod_{i=1}^n \frac{\mathbb{E}[\mathbb{1}(W_{in}\in B_i)\exp(s_nW_{in})\mid\mathcal{F}_n]}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\right],
\end{aligned} \tag{154}$$

from which it follows that

$$\mathbb{P}_{n,s_n}[W_{1n}\in B_1,\dots,W_{nn}\in B_n\mid\mathcal{F}_n] = \prod_{i=1}^n \frac{\mathbb{E}[\mathbb{1}(W_{in}\in B_i)\exp(s_nW_{in})\mid\mathcal{F}_n]}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}$$

This verifies the claim that under $\mathbb{P}_{n,s_n}$, $(W_{1n},\dots,W_{nn})$ are still independent conditionally on $\mathcal{F}_n$. Furthermore, this shows that the marginal distribution of each W_{in} is exponentially tilted by s_n, conditionally on $\mathcal{F}_n$. From this, we can derive the conditional mean and variance of W_{in} under the measure $\mathbb{P}_{n,s_n}$. We write

$$\mathbb{E}_{n,s_n}[W_{in}\mid\mathcal{F}_n] = \mathbb{E}\left[W_{in}\prod_{i=1}^n \frac{\exp(s_nW_{in})}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}\mid\mathcal{F}_n\right] = \frac{\mathbb{E}\left[W_{in}\exp(s_nW_{in})\mid\mathcal{F}_n\right]}{\mathbb{E}[\exp(s_nW_{in})\mid\mathcal{F}_n]}.$$

Then by Lemma 45, we have,

$$\mathbb{P}\left[\mathcal{T}\right]=1,\ \mathcal{T}\equiv\left\{K'_{in}(s)=\mathbb{E}_{in,s}[W_{in}|\mathcal{F}_n],\ \forall s\in(-\varepsilon,\varepsilon)\right\}.$$

Then we know $\forall\omega\in\mathcal{T}\cap\{|s_n|<\varepsilon\}$,

$$\begin{aligned}K'_{in}(s_n)(\omega)=\mathbb{E}_{in,s_n}[W_{in}|\mathcal{F}_n](\omega)&=\int x\frac{\exp(s_nx)}{\int\exp(s_nx)\mathrm{d}\kappa_{in}(\omega,x)}\mathrm{d}\kappa_{in}(\omega,x)\\&=\mathbb{E}_{n,s_n}[W_{in}|\mathcal{F}_n](\omega),\end{aligned}$$

so that by the assumption $\mathbb{P}[s_n\in(-\varepsilon,\varepsilon)]=1$,

$$\mathbb{P}\left[K'_{in}(s_n)=\mathbb{E}_{n,s_n}[W_{in}|\mathcal{F}_n]\right]=1.$$

Similarly, we have

$$\mathbb{P}\left[K''_{in}(s_n)=\mathrm{Var}_{n,s_n}[W_{in}|\mathcal{F}_n]\right]=1.$$

□

I.8 Proof of Lemma 29

Proof of Lemma 29. Define

$$F_n(t,\omega)\equiv\mathbb{P}\left[\frac{1}{S_n\sqrt{n}}\sum_{i=1}^n(W_{in}-\mathbb{E}[W_{in}|\mathcal{F}_n])\le t|\mathcal{F}_n\right](\omega)$$

and

$$S_n(\omega)\equiv\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(W_{in}-\mathbb{E}[W_{in}|\mathcal{F}_n])^2|\mathcal{F}_n](\omega)\right)^{1/2}.$$

We prove the result by recalling the notion of regular conditional distribution defined in Appendix A.4. Define $W_n\equiv(W_{1n},\ldots,W_{nn})$. Suppose $\kappa_{W_n,\mathcal{F}_n}$ is a regular conditional distribution of W_n given $\mathcal{F}_n$. Then for every $\omega\in\Omega$, we know $\kappa_{W_n,\mathcal{F}_n}(\omega,\cdot)$ is a probability measure. We draw $(\widetilde{W}_{1n}(\omega),\ldots,\widetilde{W}_{nn}(\omega))\sim\kappa_{W_n,\mathcal{F}_n}(\omega,\cdot)$. Define

$$\widetilde{S}_n(\omega)\equiv\left(\frac{1}{n}\sum_{i=1}^n\mathbb{E}[(\widetilde{W}_{in}(\omega)-\mathbb{E}[\widetilde{W}_{in}(\omega)])^2]\right)^{1/2},\ \mathcal{D}_n\equiv\left\{S_n>0\right\}.$$

In order to apply Lemma 20 to almost every $\omega\in\mathcal{D}_n\subset\Omega$, we need to verify that for those ω it is true that $\forall i\in\{1,\ldots,n\}$,

$$\sum_{i=1}^n\mathbb{E}\left[\frac{(\widetilde{W}_{in}(w)-\mathbb{E}[\widetilde{W}_{in}(w)])^2}{\widetilde{S}_n^2(\omega)}\right]=1,\ \mathbb{E}\left[\frac{\widetilde{W}_{in}(w)-\mathbb{E}[\widetilde{W}_{in}(w)]}{\widetilde{S}_n(\omega)}\right]=0\tag{155}$$

and $\widetilde{S}_n(\omega)>0$. Both claims are true by applying Lemma 4, such that for almost every $\omega\in\Omega$, we have for any positive integer p,

$$\mathbb{E}[W_{in}^p|\mathcal{F}_n](\omega)=\mathbb{E}[\widetilde{W}_{in}^p(\omega)],\ \mathbb{E}[|W_{in}|^p|\mathcal{F}_n](\omega)=\mathbb{E}[|\widetilde{W}_{in}(\omega)|^p],\ \forall i\in\{1,\ldots,n\},\ n\ge1.$$

Together with the assumption imposed in the lemma, we know conditions in (155) are satisfied for any $\omega \in \mathcal{D}_n \cap \mathcal{N}^c$, where $\mathcal{N}$ is a null set with probability measure 0. Then we apply Lemma 20 to obtain that for any fixed $t \in \mathbb{R}$ there exists a universal constant, that is independent of ω, such that $\forall \omega \in \mathcal{D}_n \cap \mathcal{N}^c$

$$\left| \mathbb{P}\left[\frac{\sum_{i=1}^n (\widetilde{W}_{in}(\omega) - \mathbb{E}[\widetilde{W}_{in}(\omega)])}{\widetilde{S}_n(\omega)\sqrt{n}} \leq t \right] - \Phi(t) \right| \leq C \frac{\sum_{i=1}^n \mathbb{E}[|\widetilde{W}_{in}(\omega) - \mathbb{E}[\widetilde{W}_{in}(\omega)]|^3]}{\widetilde{S}_n^3(\omega) n^{3/2}}$$

Then fixing any $t \in \mathbb{R}$, we again appy Lemma 4 such that for almost every $\omega \in \mathcal{C}_n \cap \mathcal{D}_n$,

$$|F_n(t,\omega) - \Phi(t)| \leq C \frac{\sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^3 | \mathcal{F}_n](\omega)}{S_n(\omega) n^{3/2}}.$$

Fix $k \in \mathbb{N}$. By the continuity of the normal CDF, there exists points $-\infty = x_0 < x_1 \cdots < x_k = \infty$ with $\Phi(x_i) = i/k$. By monotonicity, we have, for $x_{i-1} \leq t \leq x_i$,

$$F_n(t,\omega) - \Phi(t) \leq F_n(x_i,\omega) - \Phi(x_{i-1}) = F_n(x_i,\omega) - \Phi(x_i) + \frac{1}{k}$$

and

$$F_n(t,\omega) - \Phi(t) \geq F_n(x_{i-1},\omega) - \Phi(x_i) = F_n(x_{i-1},\omega) - \Phi(x_{i-1}) - \frac{1}{k}.$$

Thus for fixed $x \in \mathbb{R}$, we can bound for almost every $\omega \in \mathcal{D}_n$ that

$$\begin{aligned} |F_n(t,\omega) - \Phi(t)| &\leq \sup_i |F_n(x_i,\omega) - \Phi(x_i)| + \frac{1}{k} \\ &\leq C \frac{\sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^3 | \mathcal{F}_n](\omega)}{S_n(\omega) n^{3/2}} + \frac{1}{k}. \end{aligned}$$

Then taking the supremum on $t \in \mathbb{R}$, we have almost every $\omega \in \mathcal{D}_n$,

$$\sup_{t \in \mathbb{R}} |F_n(t,\omega) - F(t)| \leq C \frac{\sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^3 | \mathcal{F}_n](\omega)}{S_n^3(\omega) n^{3/2}} + \frac{1}{k}.$$

Letting k go to infinity, we have

$$\mathbb{1}(\omega \in \mathcal{D}_n) \sup_{t \in \mathbb{R}} |F_n(t,\omega) - F(t)| \leq C \mathbb{1}(\omega \in \mathcal{D}_n) \frac{\sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^3 | \mathcal{F}_n](\omega)}{S_n^3(\omega) n^{3/2}}. \tag{156}$$

Now we decompose

$$\begin{aligned} &\sqrt{n} \sup_{t \in \mathbb{R}} |F_n(t,\omega) - F(t)| \\ &= \sqrt{n} \mathbb{1}(\omega \notin \mathcal{D}_n) \sup_{t \in \mathbb{R}} |F_n(t,\omega) - F(t)| + \sqrt{n} \mathbb{1}(\omega \in \mathcal{D}_n) \sup_{t \in \mathbb{R}} |F_n(t,\omega) - F(t)|. \end{aligned}$$

We first show that $\mathbb{1}(\omega \notin \mathcal{D}_n) = o_{\mathbb{P}}(1)$. We only need to prove

$$\mathbb{P}[\mathcal{D}_n^c] = \mathbb{P}[S_n \leq 0] \to 0.$$

This is obvious since $S_n = \Omega_{\mathbb{P}}(1)$. Then we have $\mathbb{1}(\omega \notin \mathcal{D}_n) = o_{\mathbb{P}}(1)$ such that

$$\sqrt{n}\mathbb{1}(\omega \notin \mathcal{D}_n) \sup_{t\in\mathbb{R}} |F_n(t,\omega) - F(t)| = o_{\mathbb{P}}(1). \tag{157}$$

By (156) and

$$\frac{\sum_{i=1}^n \mathbb{E}[|W_{in} - \mathbb{E}[W_{in}|\mathcal{F}_n]|^3|\mathcal{F}_n]}{n} = O_{\mathbb{P}}(1), S_n = \Omega_{\mathbb{P}}(1),$$

we know there exists $M > 0$ such that for any $m \geq M$, we have

$$\mathbb{P}\left[\sqrt{n}\mathbb{1}(\omega \in \mathcal{D}_n) \sup_{t\in\mathbb{R}} |F_n(t,\omega) - F(t)| > m\right] \to 0.$$

This, together with (157), implies for any $m \geq M$ we have

$$\mathbb{P}\left[\sqrt{n} \sup_{t\in\mathbb{R}} |F_n(t,\omega) - F(t)| > m\right] \to 0.$$

□

I.9 Proof of Lemma 30

The statement (122) can be verified using a Taylor expansion, guaranteed by Lemma 5, of $K_n''(s)$ around $s = 0$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \mathrm{Var}_{n,\hat{s}_n}[W_{in} \mid \mathcal{F}_n] &= K_n''(\hat{s}_n) \\ &= K_n''(0) + \hat{s}_n K_n^{(3)}(0) + \frac{1}{2}\hat{s}_n^2 K_n^{(4)}(\tilde{s}_n) \\ &= \Omega_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) + o_{\mathbb{P}}(1) \\ &= \Omega_{\mathbb{P}}(1),\end{aligned}$$

where $|\tilde{s}_n| \leq |\hat{s}_n|$. The statement $K_n''(0) = \Omega_{\mathbb{P}}(1)$ is by assumption (141), the statement $\hat{s}_n K_n^{(3)}(0)$ is by the convergence $\hat{s}_n \xrightarrow{\mathbb{P}} 0$ (109) and the finiteness of $K_n^{(3)}(0)$ (39), and the statement $\hat{s}_n^2 K_n^{(4)}(\tilde{s}_n) = o_{\mathbb{P}}(1)$ is by the convergence $\hat{s}_n \xrightarrow{\mathbb{P}} 0$ and the assumption (140).

To verify the moment condition (123), by Lemma 15 with $p = 3, q = 4$, it suffices to verify a stronger fourth moment statement

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{n,\hat{s}_n}[(W_{in} - \mathbb{E}_{n,\hat{s}_n}[W_{in}])^4 \mid \mathcal{F}_n] = O_{\mathbb{P}_{n,\hat{s}_n}}(1). \tag{158}$$

To this end, we combine an expression for the fourth central moment of W_{in} in terms of the second and fourth cumulants and the assumptions (138) and (140):

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}_{n,\hat{s}_n}[(W_{in} - \mathbb{E}_{n,\hat{s}_n}[W_{in}])^4 \mid \mathcal{F}_n] = \frac{1}{n}\sum_{i=1}^n \left\{K_{in}^{(4)}(\hat{s}_n) + 3(K_{in}''(\hat{s}_n))^2\right\} = O_{\mathbb{P}}(1).$$

I.10 Proof of Lemma 31

For any $A_n \in \mathcal{F}_n$, we have

$$\begin{aligned}
&\mathbb{E}\left[\mathbb{1}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right)\mathbb{1}(A_n)\right] \\
&= \mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right)\mathbb{1}(A_n)\frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}}\right] \\
&= \mathbb{E}_{n,\hat{s}_n}\left[\mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right)\mathbb{1}(A_n)\frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}} \mid \mathcal{F}_n\right]\right] \\
&= \mathbb{E}_{n,\hat{s}_n}\left[\mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right)\frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}} \mid \mathcal{F}_n\right]\mathbb{1}(A_n)\right] \\
&= \mathbb{E}\left[\mathbb{E}_{n,\hat{s}_n}\left[\mathbb{1}\left(\frac{1}{n}\sum_{i=1}^{n} W_{in} \geq w_n\right)\frac{d\mathbb{P}}{d\mathbb{P}_{n,\hat{s}_n}} \mid \mathcal{F}_n\right]\mathbb{1}(A_n)\right].
\end{aligned}$$

The equality is due to Lemma 28, since the random variable inside the expectation is measurable with respect to $\mathcal{F}_n$.

I.11 Proof of Lemma 32

Proof of Lemma 32. Recall the notion of regular conditional distribution introduced in Appendix A.4. We know $Z_n|\mathcal{F}_n$ must admit the regular conditional distribution $\kappa_n(\omega, B)$ for $B \in \mathcal{B}(\mathbb{R}^n)$. We define $F_n(\cdot, \omega)$ to be the CDF of $Z_n|\mathcal{F}_n$ for the probability measure $\kappa_n(\omega, \cdot)$. We apply Lemma 4 and the integration by parts formula to obtain, for almost every $\omega \in \Omega$,

$$\begin{aligned}
&\mathbb{1}(\lambda_n \geq 0)\mathbb{E}\left[\mathbb{1}(Z_n \geq 0)\exp\left(-\lambda_n Z_n\right) \mid \mathcal{F}_n\right](\omega) \\
&= \mathbb{1}(\lambda_n \geq 0)\int_0^\infty \exp(-\lambda_n z)\mathrm{d}F_n(z,\omega) \\
&= -\mathbb{1}(\lambda_n \geq 0)F_n(0,\omega) + \mathbb{1}(\lambda_n \geq 0)\int_0^\infty \lambda_n \exp(-\lambda_n z)F_n(z,\omega)\mathrm{d}z. \qquad (159)
\end{aligned}$$

Similarly, apply integration by parts so that we have

$$\mathbb{1}(\lambda_n \geq 0)\int_0^\infty \exp(-\lambda_n z)\phi(z)\mathrm{d}z = \mathbb{1}(\lambda_n \geq 0)\left(-\Phi(0) + \int_0^\infty \lambda_n \exp(-\lambda_n z)\Phi(z)\mathrm{d}z\right). \qquad (160)$$

Then combining (159) and (160), we can bound

$$\begin{aligned}
&\mathbb{1}(\lambda_n \geq 0) \left| \mathbb{E}\left[\mathbb{1}(Z_n \geq 0) \exp\left(-\lambda_n Z_n\right) \mid \mathcal{F}_n\right](\omega) - \int_0^\infty \exp(-\lambda_n z)\phi(z)dz \right| \\
&= \mathbb{1}(\lambda_n \geq 0) \left| \Phi(0) - F_n(0,\omega) + \int_0^\infty \lambda_n \exp(-\lambda_n z)\left(F_n(z,\omega) - \Phi(z)\right) \mathrm{d}z \right| \\
&\leq \mathbb{1}(\lambda_n \geq 0) \sup_{z\geq 0} |F_n(z,\omega) - \Phi(z)| \left(1 + \int_0^\infty \lambda_n \exp(-\lambda_n z)\mathrm{d}z\right) \\
&= 2\mathbb{1}(\lambda_n \geq 0) \sup_{z\geq 0} |F_n(z,\omega) - \Phi(z)| \\
&\leq 2\mathbb{1}(\lambda_n \geq 0) \sup_{z\in\mathbb{R}} |F_n(z,\omega) - \Phi(z)| \\
&= 2\mathbb{1}(\lambda_n \geq 0) \sup_{z\in\mathbb{R}} |\mathbb{P}\left[Z_n \leq z|\mathcal{F}_n\right](\omega) - \Phi(z)|
\end{aligned}$$

almost surely. For the last equality, we use Lemma 4 together with the density argument to prove the equality. Indeed, fixing any $k \in \mathbb{N}$, by the continuity of the normal CDF, there exists points $-\infty = x_0 < x_1 \cdots < x_k = \infty$ with $\Phi(x_i) = i/k$. By monotonicity, we have for $x_{i-1} \leq t \leq x_i$

$$F_n(t,\omega) - \Phi(t) \leq F_n(x_i,\omega) - \Phi(x_{i-1}) = \mathbb{P}[Z_n \leq x_i|\mathcal{F}_n](\omega) - \Phi(x_i) + \frac{1}{k}$$

and

$$F_n(t,\omega) - \Phi(t) \geq F_n(x_{i-1},\omega) - \Phi(x_i) = \mathbb{P}[Z_n \leq x_{i-1}|\mathcal{F}_n](\omega) - \Phi(x_{i-1}) - \frac{1}{k}$$

for almost every $\omega \in \Omega$. Then we have for almost every $\omega \in \Omega$

$$\begin{aligned}
|F_n(t,\omega) - \Phi(t)| &\leq \sup_i |\mathbb{P}[Z_n \leq x_i|\mathcal{F}_n](\omega) - \Phi(x_i)| + \frac{1}{k} \\
&\leq \sup_{t\in\mathbb{R}} |\mathbb{P}[Z_n \leq t|\mathcal{F}_n](\omega) - \Phi(t)| + \frac{1}{k}.
\end{aligned}$$

Therefore by the arbitrary choice of k so that we have

$$\sup_{t\in\mathbb{R}} |F_n(t,\omega) - \Phi(t)| \leq \sup_{t\in\mathbb{R}} |\mathbb{P}[Z_n \leq t|\mathcal{F}_n](\omega) - \Phi(t)|$$

almost surely. By interchanging the $F_n(t,\omega)$ and $\mathbb{P}[Z_n \leq t|\mathcal{F}_n](\omega)$, we have shown the desired result. □

I.12 Proof of Lemma 33

Proof of statement (132). We consider the events $r_n = 0$ and $r_n > 0$ separately. First, define $\mathcal{U}_n \equiv \{U_n \neq 0\}$.

On the event $r_n = 0$: We further divide this case into two cases.

1. When $\lambda_n = 0$: this implies $|U_n| = 1/2 > 0$;
2. When $\lambda_n \neq 0$: this implies $|U_n| = \infty$.

This implies $\mathcal{U}_n$ happens.

On the event $r_n > 0$: By (Sign4) condition, this implies $\lambda_n \geq 0$. We divide the case to $\lambda_n > 0$ and $\lambda_n = 0$.

1. When $\lambda_n = 0$: this implies $|U_n| = \infty$;
2. When $\lambda_n > 0$: we discuss when $r_n - \lambda_n > 0, r_n - \lambda_n < 0$ and $r_n - \lambda_n = 0$.

When $r_n - \lambda_n \geq 0$: By the formula of U_n, we have

$$\begin{aligned} U_n &= \exp\left(\frac{r_n^2}{2}\right)(1 - \Phi(r_n)) + \frac{1}{\sqrt{2\pi}}\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\} \\ &\geq \exp\left(\frac{r_n^2}{2}\right)(1 - \Phi(r_n)). \end{aligned}$$

By (Finite), we know $r_n \in (-\infty, \infty)$ almost surely, so that $U_n > 0$ almost surely.

When $r_n - \lambda_n < 0$: In order to proceed the proof, we present a lemma to relate the U_n with the Gaussian integral estimate via integration by parts.

Lemma 42. *Suppose* (Finite) *condition is true. Define*

$$R_n \equiv \int_{r_n}^{\lambda_n} y \exp(y^2/2)(1 - \Phi(y)) - \frac{1 - y^{-2}}{\sqrt{2\pi}} \mathrm{d}y. \tag{161}$$

Recall the definition of U_n as in (129)*. If $\lambda_n, r_n \neq 0$ almost surely, then we have*

$$R_n = \int_0^{\infty} \exp(-\lambda_n y)\phi(y)\mathrm{d}y - U_n, \text{ almost surely.}$$

Since in this case, $\lambda_n > r_n > 0$ and (Finite) is assumed in the lemma statement, then by Lemma 42 we have

$$U_n = \int_0^{\infty} \exp(-\lambda_n y)\phi(y)\mathrm{d}y - R_n, \text{ almost surely.}$$

By Lemma 17, we can write

$$R_n = \int_{r_n}^{\lambda_n} \left(-y \int_y^{\infty} \frac{\phi(t)}{3t^4}\mathrm{d}t\right)\mathrm{d}y < 0,$$

which implies $U_n > 0$. Thus $\mathcal{U}_n$ happens.

On the event $r_n < 0$: By (Sign4) condition, we know $\lambda_n \leq 0$. We divide the case to $\lambda_n < 0$ and $\lambda_n = 0$.

1. When $\lambda_n < 0$: we can lower bound

$$U_n = \exp\left(\frac{r_n^2}{2}\right)(1 - \Phi(r_n)) + \frac{1}{\sqrt{2\pi}}\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\} \geq \frac{1}{2} - \frac{1}{\sqrt{2\pi}}\left|\frac{1}{\lambda_n} - \frac{1}{r_n}\right|$$

By (Rate1), we have

$$\frac{1}{\sqrt{2\pi}}\left|\frac{1}{\lambda_n}-\frac{1}{r_n}\right| = o_{\mathbb{P}}(1).$$

Thus we have

$$\mathbb{P}\left[\mathcal{U}_n^c \text{ and } r_n < 0 \text{ and } \lambda_n < 0\right] \leq \mathbb{P}\left[U_n < \frac{1}{4} \text{ and } r_n < 0\right] \to 0.$$

2. When $\lambda_n = 0$: we know $|U_n| = \infty$. Thus $\mathcal{U}_n$ happens. This completes the proof.

□

Proof of statement (133). We write

$$\mathbb{1}(r_n \geq 0)\frac{1}{\sqrt{n}U_n} = \mathbb{1}(r_n \geq 1)\frac{1}{r_n U_n}\frac{r_n}{\sqrt{n}} + \mathbb{1}(r_n \in [0,1))\frac{1}{\sqrt{n}U_n}.$$

Now we present an auxiliary lemma.

Lemma 43 (Convergence rate of $1/U_n$)**.** *Suppose* (Rate1) *and* (Rate2) *hold. Then we have*

$$\frac{\mathbb{1}(r_n \geq 1)}{r_n U_n} = O_{\mathbb{P}}(1) \tag{162}$$

$$\mathbb{1}(r_n \in [0,1))\frac{1}{U_n} = O_{\mathbb{P}}(1). \tag{163}$$

Intuition of Lemma 43: The intuition behind this is when r_n is small, we expect $|U_n|$, is lower bounded with high probability since $1/\lambda_n - 1/r_n = o_{\mathbb{P}}(1)$ and thus the dominant term is $\exp(r_n^2/2)(1-\Phi(r_n))$, which is lower bounded when r_n is small. When r_n is large, $|U_n|$ will go to zero but with a rate that is slower than $1/r_n$. The latter case needs a finer analysis with (Rate1) and (Rate2) conditions involved.

Then by Lemma 43, we know

$$\mathbb{1}(r_n \geq 1)\frac{1}{r_n U_n} = O_{\mathbb{P}}(1),\ \ \mathbb{1}(r_n \in [0,1))\frac{1}{U_n} = O_{\mathbb{P}}(1).$$

Since $r_n/\sqrt{n} = o_{\mathbb{P}}(1)$, we conclude

$$\mathbb{1}(r_n \geq 0)\frac{1}{\sqrt{n}U_n} = o_{\mathbb{P}}(1).$$

This completes the proof. □

I.13 Proof of Lemma 34

Proof of Lemma 34. Define

$$O_n \equiv \frac{|r_n - \lambda_n|}{\sqrt{2\pi}} \left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2} \right).$$

We first present an auxiliary lemma.

Lemma 44 (Upper bound of Gaussian integral)**.** *Under conditions* (Finite) *and* (Sign4)*, the following inequality is true almost surely:*

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \left| \frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} - 1 \right| \leq \mathbb{1}(r_n > 0, \lambda_n > 0) \left| \frac{1}{U_n} \right| \cdot O_n.$$

By Lemma 44, we can bound

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \left| \frac{\int_0^\infty \exp(-\lambda_n z)\phi(z)dz}{U_n} - 1 \right| \leq \left| \mathbb{1}(r_n > 0, \lambda_n > 0) \frac{O_n}{U_n} \right|$$

almost surely. Then we can further decompose

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \frac{O_n}{U_n} = \frac{\mathbb{1}(r_n \geq 1, \lambda_n > 0)}{r_n U_n} \cdot r_n O_n + \mathbb{1}(r_n \in (0,1), \lambda_n > 0) \frac{O_n}{U_n}$$

By Lemma 43, we know

$$\mathbb{1}(r_n \geq 1) \frac{1}{r_n U_n} = O_{\mathbb{P}}(1), \;\; \mathbb{1}(r_n \in (0,1)) \frac{1}{U_n} = O_{\mathbb{P}}(1).$$

Thus it suffices to show

$$r_n O_n = o_{\mathbb{P}}(1) \tag{164}$$

and

$$\mathbb{1}(r_n > 0, \lambda_n > 0) O_n = o_{\mathbb{P}}(1). \tag{165}$$

Proof of (164)**:** We compute

$$r_n O_n = \frac{1}{\sqrt{2\pi}} \left| 1 - \frac{\lambda_n}{r_n} \right| \cdot \left| 1 + \frac{r_n^2}{\lambda_n^2} \right|.$$

Thus by (Rate2) we know

$$\left| \frac{\lambda_n}{r_n} - 1 \right| = o_{\mathbb{P}}(1), \;\; \frac{r_n^2}{\lambda_n^2} = O_{\mathbb{P}}(1).$$

Thus we have $\mathbb{1}(r_n \geq 1, \lambda_n > 0) r_n O_n = o_{\mathbb{P}}(1)$.

Proof of (165): We can write

$$O_n = \frac{|r_n - \lambda_n|}{\sqrt{2\pi}} \left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2} \right) = \frac{1}{\sqrt{2\pi}} \left| \left(\frac{\lambda_n}{r_n} - 1 \right) \frac{1}{r_n} \right| \cdot \left| 1 + \frac{r_n^2}{\lambda_n^2} \right|.$$

Then by (Rate3), we know

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \left| \left(\frac{\lambda_n}{r_n} - 1 \right) \frac{1}{r_n} \right| = o_{\mathbb{P}}(1)$$

and by (Rate2), we have $r_n^2/\lambda_n^2 = O_{\mathbb{P}}(1)$. Thus we have $O_n = o_{\mathbb{P}}(1)$. □

I.14 Proof of Lemma 35

Proof of Lemma 35. By the rate condition (Rate2), we know

$$1 + \frac{\lambda_n^2}{r_n^2} = O_{\mathbb{P}}(1).$$

Thus we only need to show

$$\mathbb{1}(\lambda_n \neq 0) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

We decompose the magnitude of $|\lambda_n|$ to two parts: $|\lambda_n| > 1$ and $|\lambda_n| \in (0, 1]$. It suffices to prove

$$\mathbb{1}(|\lambda_n| \in (0,1]) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1), \;\; \mathbb{1}(|\lambda_n| > 1) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1). \tag{166}$$

For the first term in (166), we know $h(x)$ is unformly lower bounded for $x \in [-1, 1]$ so that $h(\lambda_n)$ is uniformly lower bounded for $|\lambda_n| \in (0, 1]$. Then by the rate condition (Rate4), we know

$$\mathbb{1}(|\lambda_n| \in (0,1]) \frac{1 - \frac{r_n}{\lambda_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1).$$

For the second term in (166), we have by Lemma 18 that for $|\lambda_n| > 1$,

$$\begin{aligned} |\lambda_n h(\lambda_n)| = |\lambda_n| \exp(\lambda_n^2/2)(1 - \Phi(\lambda_n)) &\geq |\lambda_n| \exp(\lambda_n^2/2)(1 - \Phi(|\lambda_n|)) \\ &\geq \frac{1}{\sqrt{2\pi}} \frac{\lambda_n^2}{\lambda_n^2 + 1} > \frac{1}{2\sqrt{2\pi}}. \end{aligned}$$

Then by the rate condition (Rate2), we know

$$\mathbb{1}(|\lambda_n| > 1) \frac{|1 - \frac{r_n}{\lambda_n}|}{|\lambda_n h(\lambda_n)|} \leq 2\sqrt{2\pi} \mathbb{1}(|\lambda_n| > 1) \left| 1 - \frac{r_n}{\lambda_n} \right| = o_{\mathbb{P}}(1).$$

□

I.15 Proof of Lemma 36

Proof of Lemma 36. We decompose the magnitude of $|\lambda_n|$ to two parts $|\lambda_n| > 1$ and $|\lambda_n| \in (0, 1]$. It suffices to prove

$$\mathbb{1}(|\lambda_n| \in (0,1]) \frac{1 - \frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1), \ \ \mathbb{1}(|\lambda_n| > 1) \frac{1 - \frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = o_{\mathbb{P}}(1). \tag{167}$$

For the first term in (167), we know $h(x)$ is uniformly bounded for $x \in [-1, 1]$ so that $h(\lambda_n)$ is uniformly bounded for $|\lambda_n| \in (0, 1]$. Then by the rate condition (Rate1), we know

$$\mathbb{1}(|\lambda_n| \in (0,1]) \frac{1 - \frac{\lambda_n}{r_n}}{\lambda_n h(\lambda_n)} = \mathbb{1}(|\lambda_n| \in (0,1]) \frac{\frac{1}{\lambda_n} - \frac{1}{r_n}}{h(\lambda_n)} = o_{\mathbb{P}}(1).$$

For the second term in (167), we have by Lemma 18 that for $|\lambda_n| > 1$,

$$\begin{aligned} |\lambda_n h(\lambda_n)| = |\lambda_n| \exp(\lambda_n^2/2)(1 - \Phi(\lambda_n)) &\geq |\lambda_n| \exp(\lambda_n^2/2)(1 - \Phi(|\lambda_n|)) \\ &\geq \frac{1}{\sqrt{2\pi}} \frac{\lambda_n^2}{\lambda_n^2 + 1} > \frac{1}{2\sqrt{2\pi}}. \end{aligned}$$

Then by the rate condition (Rate2), we know

$$\mathbb{1}(|\lambda_n| > 1) \frac{|1 - \frac{\lambda_n}{r_n}|}{|\lambda_n h(\lambda_n)|} \leq 2\sqrt{2\pi} \mathbb{1}(|\lambda_n| > 1) \left| 1 - \frac{\lambda_n}{r_n} \right| = o_{\mathbb{P}}(1).$$

□

I.16 Proof of Lemma 37

Proof of Lemma 37. The following lemma states how the derivatives of $K_{in}(s)$ are related to the conditional moments of $W_{in}|\mathcal{F}_n$ under measure $\kappa_{in,s}$.

Lemma 45. *On the event $\mathcal{A}$ as in Lemma 5, we have*

$$K'_{in}(s) = \mathbb{E}_{in,s}[W_{in}|\mathcal{F}_n], \ \ \forall s \in (-\varepsilon, \varepsilon), \tag{168}$$

$$K''_{in}(s) = \mathrm{Var}_{in,s}[W_{in}|\mathcal{F}_n], \ \ \forall s \in (-\varepsilon, \varepsilon), \tag{169}$$

$$K^{(4)}_{in}(s) = \mathbb{E}_{in,s}[(W_{in} - \mathbb{E}_{in,s}[W_{in}|\mathcal{F}_n])^4|\mathcal{F}_n] - 3\mathrm{Var}^2_{in,s}[W_{in}|\mathcal{F}_n], \ \ \forall s \in (-\varepsilon, \varepsilon). \tag{170}$$

We first show with Lemma 45, in order to show condition (138)-(140), it suffices to show there exists $\varepsilon > 0$ such that $\mathbb{P}[\mathcal{A}] = 1$ and for the given $\varepsilon > 0$,

$$\sup_{s \in (-\varepsilon, \varepsilon)} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{in,s}[W^4_{in}|\mathcal{F}_n] = O_{\mathbb{P}}(1). \tag{171}$$

Suppose $\mathbb{P}[\mathcal{A}] = 1$ and the assumption (171) holds. Now we verify condition (138)-(140) subsequently.

Verification of condition (138): By conclusion (169), Jensen's inequality and statement (171), we have

$$\begin{aligned}
\sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}(K_{in}''(s))^2 &\leq \sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}(\mathbb{E}_{in,s}[W_{in}^2|\mathcal{F}_n])^2 \\
&= \sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}\left(\int x^2 \mathrm{d}\kappa_{in,s}(\omega,x)\right)^2 \\
&\leq \sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}\int x^4 \mathrm{d}\kappa_{in,s}(\omega,x) \\
&= \sup_{s\in(-\varepsilon,\varepsilon)} \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{in,s}[W_{in}^4|\mathcal{F}_n] \\
&= O_{\mathbb{P}}(1).
\end{aligned}$$

Verification of condition (139): It suffices to prove

$$\frac{1}{n}\sum_{i=1}^{n}K_{in}'''(0) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[W_{in}^3|\mathcal{F}_n] = O_{\mathbb{P}}(1).$$

By Lemma 15 with $p=3, q=4$, we can bound

$$\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[W_{in}^3|\mathcal{F}_n]\right| \leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|W_{in}|^3|\mathcal{F}_n] \leq \left(\frac{\sum_{i=1}^{n}\mathbb{E}[W_{in}^4|\mathcal{F}_n]}{n}\right)^{3/4}.$$

Then by statement (171), we have

$$\left|\frac{1}{n}\sum_{i=1}^{n}K_{in}'''(0)\right| \leq \left(\frac{\sum_{i=1}^{n}\mathbb{E}[W_{in}^4|\mathcal{F}_n]}{n}\right)^{3/4} = O_{\mathbb{P}}(1).$$

Verification of condition (140)**:** By conclusion (170) and (169),

$$
\begin{aligned}
&\sup_{s\in(-\varepsilon,\varepsilon)}\left|\frac{1}{n}\sum_{i=1}^{n}K_{in}''''(s)\right| \\
&\leq \sup_{s\in(-\varepsilon,\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{in,s}[(W_{in}-\mathbb{E}_{in,s}[W_{in}|\mathcal{F}_n])^4|\mathcal{F}_n]+\sup_{s\in(-\varepsilon,\varepsilon)}\frac{3}{n}\sum_{i=1}^{n}(K_{in}''(s))^2 \\
&= \sup_{s\in(-\varepsilon,\varepsilon)}\frac{1}{n}\sum_{i=1}^{n}\int\left(x-\int x\mathrm{d}\kappa_{in,s}(\cdot,x)\right)^4\mathrm{d}\kappa_{in,s}(\cdot,x)+O_{\mathbb{P}}(1) \\
&\leq \sup_{s\in(-\varepsilon,\varepsilon)}\frac{16}{n}\sum_{i=1}^{n}\left(\int x^4\mathrm{d}\kappa_{in,s}(\cdot,x)+\left(\int x\mathrm{d}\kappa_{in,s}(\cdot,x)\right)^4\right)+O_{\mathbb{P}}(1) \\
&\leq \sup_{s\in(-\varepsilon,\varepsilon)}\frac{32}{n}\sum_{i=1}^{n}\int x^4\mathrm{d}\kappa_{in,s}(\cdot,x)+O_{\mathbb{P}}(1) \\
&= \sup_{s\in(-\varepsilon,\varepsilon)}\frac{32}{n}\sum_{i=1}^{n}\mathbb{E}_{in,s}[W_{in}^4|\mathcal{F}_n]+O_{\mathbb{P}}(1) \\
&= O_{\mathbb{P}}(1)
\end{aligned}
$$

where the third inequality is due to power inequality $(|a|+|b|)^p \leq 2^p(|a|^p+|b|^p)$ and the proved condition (138), the fourth inequality is due to Jensen's inequality and the last equality is due to statement (171). Now we show there exists $\varepsilon > 0$ such that statement (171) holds for both cases.

Case 1: CSE distribution Consider the power-series expansion

$$
\begin{aligned}
\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] &= \int \exp(sx)\mathrm{d}\kappa_{in}(\omega,x) \\
&= \int\left(1+\sum_{k=1}^{\infty}\frac{s^k}{k!}x^k\right)\mathrm{d}\kappa_{in}(\omega,x) \\
&= 1+\sum_{k=1}^{\infty}\frac{s^k}{k!}\int x^k\mathrm{d}\kappa_{in}(\omega,x) \\
&= 1+\sum_{k=2}^{\infty}\frac{s^k}{k!}\mathbb{E}[W_{in}^k|\mathcal{F}_n] \qquad (172)
\end{aligned}
$$

where the second last inequality is due to Fubini's theorem and the last inequality is due to the definition of conditional expectation (32). We first can bound using Lemma 21 and Assumption 1 that

$$
\mathbb{P}\left[\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] \leq \exp(\lambda_n s^2), \forall s\in(-\beta/4,\beta/4)\right] = 1 \qquad (173)
$$

where

$$
\lambda_n = \frac{\sqrt{6!4^6}(1+\theta_n)}{24\beta^2}+\frac{16(1+\theta_n)}{\beta^2}.
$$

Then we can bound by setting $s = \beta/16$ in the identity (172) and conclusion (173) so that

$$\mathbb{E}[W_{in}^{12}|\mathcal{F}_n] \leq \frac{12!16^{12}}{\beta^{12}}\mathbb{E}\left[\exp\left(\frac{\beta}{16}W_{in}\right)|\mathcal{F}_n\right] \leq \frac{12!16^{12}}{\beta^{12}}\exp\left(\frac{\lambda_n\beta^2}{256}\right) \tag{174}$$

almost surely. Then we can bound for $|s| < \beta/8$:

$$\begin{aligned}
\mathbb{E}_{in,s}[W_{in}^4|\mathcal{F}_n] &= \frac{\mathbb{E}[W_{in}^4\exp(sW_{in})|\mathcal{F}_n]}{\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n]} \\
&= \frac{\int x^4\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x)}{\int\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x)} \\
&\leq \int x^4\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x) \\
&\leq \left(\int x^{12}\mathrm{d}\kappa_{in}(\cdot,x)\right)^{1/3}\left(\int\exp\left(\frac{3sx}{2}\right)\mathrm{d}\kappa_{in}(\cdot,x)\right)^{2/3} \\
&= \left(\mathbb{E}[W_{in}^{12}|\mathcal{F}_n]\right)^{1/3}\left(\mathbb{E}[\exp(3sW_{in}/2)|\mathcal{F}_n]\right)^{2/3} \\
&\leq \left(\frac{12!16^{12}}{\beta^{12}}\right)^{1/3}\exp\left(\frac{\lambda_n\beta^2}{768}\right)\cdot\exp\left(\frac{3\lambda_n\beta^2}{128}\right)
\end{aligned}$$

where the third and fourth inequality is due to Jensen's inequality and Hölder's inequality, respectively and the last inequality is due to bound (174) and bound (173). By the assumption $\lambda_n = O_{\mathbb{P}}(1)$, we have

$$\sup_{s\in(-\varepsilon,\varepsilon)}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{in,s}[W_{in}^4|\mathcal{F}_n] \leq \left(\frac{12!16^{12}}{\beta^{12}}\right)^{1/3}\exp\left(\frac{\lambda_n\beta^2}{768}\right)\cdot\exp\left(\frac{3\lambda_n\beta^2}{128}\right) = O_{\mathbb{P}}(1).$$

Case 2: CCS distribution By Lemma 5, we know $\mathbb{P}[\mathcal{A}] = 1$ with $\varepsilon = 1$. Since $\mathbb{P}[\mathrm{Supp}(\kappa_{in}(\omega,\cdot)) \in [-\nu_{in}(\omega),\nu_{in}(\omega)]] = 1$, we can bound,

$$\begin{aligned}
\mathbb{E}_{in,s}\left[W_{in}^4|\mathcal{F}_n\right] &= \int x^4\frac{\exp(sx)}{\int\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x)}\mathrm{d}\kappa_{in}(\cdot,x) \\
&\leq \frac{\nu_{in}^4\int\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x)}{\int\exp(sx)\mathrm{d}\kappa_{in}(\cdot,x)} = \nu_{in}^4,\ \forall s\in(-\varepsilon,\varepsilon) = (-1,1)
\end{aligned}$$

almost surely. Thus we have

$$\sup_{s\in(-\varepsilon,\varepsilon)}\frac{1}{n}\sum_{i=1}^n\mathbb{E}_{in,s}\left[W_{in}^4|\mathcal{F}_n\right] \leq \frac{1}{n}\sum_{i=1}^n\nu_{in}^4 = O_{\mathbb{P}}(1).$$

□

I.17 Proof of Lemma 38

Proof of Lemma 38. By conclusion (169) in Lemma 45, we know on the event $\mathcal{A}$,

$$\frac{1}{n}\sum_{i=1}^n K_{in}''(0) = \frac{1}{n}\sum_{i=1}^n\mathbb{E}[W_{in}^2|\mathcal{F}_n].$$

By Lemma 5, we know $\mathbb{P}[\mathcal{A}] = 1$ and together with the condition (7), the claim is true. □

I.18 Proof of Lemma 39

Proof of Lemma 39. Define

$$A_{in,s} \equiv \mathbb{E}[|W_{in}|^p \exp(sW_{in})|\mathcal{F}_n].$$

Fix any $s_0 \in (-\varepsilon, \varepsilon)$ and suppose $a_0 \in (1, \varepsilon/|s_0|)$. We have by Hölder's inequality

$$A_{in,s_0} \leq \left\{\mathbb{E}\left[|W_{in}|^{\frac{pa_0}{a_0-1}}|\mathcal{F}_n\right]\right\}^{(a_0-1)/a_0} \left\{\mathbb{E}\left[\exp\left(a_0 s_0 W_{in}\right)|\mathcal{F}_n\right]\right\}^{1/a_0}.$$

We first show that on the event $\mathcal{A}$, all the conditional moments for $W_{in}|\mathcal{F}_n$ are finite almost surely.

Lemma 46. *On the event $\mathcal{A}$,*

$$\mathbb{E}[|W_{in}|^m|\mathcal{F}_n] < \infty, \ \forall i \in \{1, \ldots, n\}, \ \forall m \in \mathbb{N}.$$

By Lemma 46, on the event $\mathcal{A}$, we have $\mathbb{E}\left[|W_{in}|^{\lceil pa_0/(a_0-1)\rceil}|\mathcal{F}_n\right] < \infty$. We can show $a_0 s_0 < \varepsilon$ so that on the same event, $\mathbb{E}\left[\exp\left(a_0 s_0 W_{in}\right)|\mathcal{F}_n\right] < \infty$. Therefore we have proved on the event $\mathcal{A}$,

$$\mathbb{E}[|W_{in}|^p \exp(sW_{in})|\mathcal{F}_n] < \infty, \ \forall s \in (-\varepsilon, \varepsilon), \ \forall i \in \{1, \ldots, n\}, \ \forall n, p \in \mathbb{N}.$$

□

I.19 Proof of Lemma 40

Proof of Lemma 40. We first present several auxiliary results.

Auxiliary results:

$$\mathbb{P}\left[\hat{s}_n w_n - K_n(\hat{s}_n) \leq 0 \text{ and } \hat{s}_n \neq 0\right] \to 0; \tag{175}$$

$$\mathbb{P}[\lambda_n r_n \leq 0 \text{ and } \hat{s}_n \neq 0] \to 0; \tag{176}$$

$$r_n^2 = 2n(\hat{s}_n w_n - K_n(\hat{s}_n)) = n\hat{s}_n^2\left(K_n''(0) + \hat{s}_n O_{\mathbb{P}}(1)\right); \tag{177}$$

$$\lambda_n^2 = n\hat{s}_n^2 K_n''(\hat{s}_n) = n\hat{s}_n^2\left(K_n''(0) + \hat{s}_n O_{\mathbb{P}}(1)\right). \tag{178}$$

Proofs of Auxiliary results:

Proof of (175) Guaranteed by Lemma 5, we Taylor expand, for $s \in (-\varepsilon, \varepsilon)$,

$$\begin{aligned} K_n(s) &= K_n(0) + sK_n'(0) + \frac{1}{2}s^2 K_n''(0) + \frac{s^3}{6}K_n'''(0) + \frac{s^4}{24}K_n''''(\bar{s}) \\ &= \frac{1}{2}s^2 K_n''(0) + \frac{s^3}{6}K_n'''(0) + \frac{s^4}{24}K_n''''(\bar{s}(s)). \end{aligned} \tag{179}$$

where $\bar{s}(s) \in (-\varepsilon, \varepsilon)$ and the last equality is due to $K_n(0) = 0$ and $K_n'(0) = 0$. Similarly, we obtain for $s \in (-\varepsilon, \varepsilon)$,

$$\begin{aligned} sK_n'(s) &= sK_n'(0) + K_n''(0)s^2 + \frac{s^3}{2}K_n'''(0) + \frac{s^4}{6}K_n''''(\tilde{s}) \\ &= K_n''(0)s^2 + \frac{s^3}{2}K_n'''(0) + \frac{s^4}{6}K_n''''(\tilde{s}(s)) \end{aligned} \tag{180}$$

where $\tilde{s}(s) \in [-s, s] \subset (-\varepsilon, \varepsilon)$. Then subtracting the expansion (180) from the expansion (179) and setting $s = \hat{s}_n$ since $\hat{s}_n \in [-\varepsilon/2, \varepsilon/2]$, we get

$$\hat{s}_n K_n'(\hat{s}_n) - K_n(\hat{s}_n) = \frac{\hat{s}_n^2}{2} K_n''(0) + \frac{\hat{s}_n^3}{3} K_n'''(0) + \frac{\hat{s}_n^4}{6}\left(K_n''''(\tilde{s}(\hat{s}_n)) - \frac{1}{4} K_n''''(\bar{s}(\hat{s}_n)) \right). \tag{181}$$

Notice (181) is similar to our target but still differs. To account such difference, we consider

$$\begin{aligned}
&\hat{s}_n w_n - K_n(\hat{s}_n) \\
&= (\hat{s}_n K_n'(\hat{s}_n) - K_n(\hat{s}_n)) \mathbb{1}(K_n'(\hat{s}_n) = w_n) + (\hat{s}_n w_n - K_n(\hat{s}_n)) \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) \\
&= \frac{\hat{s}_n^2}{2} K_n''(0) - \frac{\hat{s}_n^2}{2} K_n''(0) \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) + \frac{\hat{s}_n^3}{3} K_n'''(0) \mathbb{1}(K_n'(\hat{s}_n) = w_n) \\
&\quad + \frac{\hat{s}_n^4}{6} (K_n''''(\tilde{s}(\hat{s}_n)) - K_n''''(\bar{s}(\hat{s}_n))/4) \mathbb{1}(K_n'(\hat{s}_n) = w_n) + (\hat{s}_n w_n - K_n(\hat{s}_n)) \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) \\
&\equiv \frac{\hat{s}_n^2}{2} K_n''(0) + \hat{s}_n^3 M_n
\end{aligned} \tag{182}$$

where M_n is a random variable that is $O_{\mathbb{P}}(1)$. This is true because the following claims are true:

$$\frac{\hat{s}_n^3}{3} K_n'''(0) \mathbb{1}(K_n'(\hat{s}_n) = w_n) = \hat{s}_n^3 O_{\mathbb{P}}(1) \tag{183}$$

$$\frac{\hat{s}_n^4}{6} (K_n''''(\tilde{s}(\hat{s}_n)) - K_n''''(\bar{s}(\hat{s}_n))/4) \mathbb{1}(K_n'(\hat{s}_n) = w_n) = \hat{s}_n^3 O_{\mathbb{P}}(1) \tag{184}$$

$$\frac{\hat{s}_n^2}{2} K_n''(0) \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) = \hat{s}_n^3 O_{\mathbb{P}}(1) \tag{185}$$

$$(\hat{s}_n w_n - K_n(\hat{s}_n)) \mathbb{1}(K_n'(\hat{s}_n) \neq w_n) = \hat{s}_n^3 O_{\mathbb{P}}(1). \tag{186}$$

Now we prove the claims (183)-(186). For claim (183), by condition (139), we know it is true. For claim (184), from condition (140) and $\tilde{s}(\hat{s}_n), \bar{s}(\hat{s}_n) \in (-\varepsilon, \varepsilon)$, we have

$$|K_n''''(\tilde{s}(\hat{s}_n))| = O_{\mathbb{P}}(1), |K_n''''(\bar{s}(\hat{s}_n))| = O_{\mathbb{P}}(1).$$

Then together with $\hat{s}_n = o_{\mathbb{P}}(1)$ For claim (185), we know it is true since $\mathbb{1}(K_n'(\hat{s}_n) \neq w_n) = o_{\mathbb{P}}(1)$. Similar argument applies to (186). Now define the event

$$\mathcal{P}_n \equiv \{\hat{s}_n \neq 0\} \cap \{K_n''(0) \leq -2\hat{s}_n M_n\}. \tag{187}$$

By condition (141), we know there exists $\eta > 0$ such that $\mathbb{P}[K_n''(0) > \eta] \to 1$. For such η since $M_n = O_{\mathbb{P}}(1)$ and $\hat{s}_n = o_{\mathbb{P}}(1)$ by Lemma 23, we have $\mathbb{P}[-2\hat{s}_n M_n < \eta] \to 1$. Together we conclude $\mathbb{P}[K_n''(0) \leq -2\hat{s}_n M_n] \to 0$. This implies $\mathbb{P}[\mathcal{P}_n] \to 0$. Moreover, on the event $\mathcal{P}_n$ we know $\hat{s}_n w_n - K_n(\hat{s}_n) \leq 0$ and $\hat{s}_n \neq 0$ happen. Therefore we conclude the proof.

Proof of (176) We compute

$$r_n \lambda_n = \begin{cases} |\hat{s}_n| \sqrt{n K_n''(\hat{s}_n)} \sqrt{2n(\hat{s}_n w_n - K_n(\hat{s}_n))} & \text{if } \hat{s}_n w_n - K_n(\hat{s}_n) \geq 0 \\ |\hat{s}_n| \sqrt{n K_n''(\hat{s}_n)} & \text{otherwise.} \end{cases}$$

Lemma 30 implies that $K_n''(\hat{s}_n) = \Omega_{\mathbb{P}}(1)$. By (175), we know $\mathbb{P}[\hat{s}_n w_n - K_n(\hat{s}_n) \leq 0 \text{ and } \hat{s}_n \neq 0] \to 0$. Moreover, $K_n''(\hat{s}_n) = \Omega_{\mathbb{P}}(1)$ can further imply $\mathbb{P}[K_n''(\hat{s}_n) = 0] \to 0$. Collecting all these, we reach

$$\begin{aligned}\mathbb{P}[r_n\lambda_n \leq 0 \text{ and } \hat{s}_n \neq 0] &= \mathbb{P}[r_n\lambda_n = 0 \text{ and } \hat{s}_n \neq 0] \\ &= \mathbb{P}[\{K_n''(\hat{s}_n) = 0 \text{ or } \hat{s}_n w_n - K_n(\hat{s}_n) = 0\} \text{ and } \{\hat{s}_n \neq 0\}] \\ &\leq \mathbb{P}[K_n''(\hat{s}_n) = 0 \text{ and } \hat{s}_n \neq 0] + \mathbb{P}[\hat{s}_n w_n - K_n(\hat{s}_n) = 0 \text{ and } \hat{s}_n \neq 0] \\ &\to 0.\end{aligned}$$

Proof of (177) For r_n^2, we can write, according to (182),

$$r_n^2 = 2n(\hat{s}_n w_n - K_n(\hat{s}_n)) = 2n\left(\frac{1}{2}\hat{s}_n^2 K_n''(0) + \hat{s}_n^3 M_n\right) = n\hat{s}_n^2 K_n''(0) + n\hat{s}_n^3 O_{\mathbb{P}}(1). \quad (188)$$

Proof of (178) We expand $K_n''(s)$ in the neighborhood $(-\varepsilon, \varepsilon)$:

$$K_n''(s) = K_n''(0) + sK_n'''(0) + \frac{s^2}{2}K_n''''(\dot{s}(s)),\ \dot{s}(s) \in (-\varepsilon, \varepsilon).$$

Then plugging $\hat{s}_n$ into above formula and observing $K_n''''(\dot{s}(\hat{s}_n)) = O_{\mathbb{P}}(1)$ ensured by condition (140) and $\dot{s}(\hat{s}_n) \in (-\varepsilon, \varepsilon)$, we obtain

$$\begin{aligned}\lambda_n^2 = n\hat{s}_n^2 K_n''(\hat{s}_n) &= n\hat{s}_n^2 K_n''(0) + n\hat{s}_n^3 K_n'''(0) + n\frac{\hat{s}_n^4}{2}K_n''''(\dot{s}(\hat{s}_n)) \\ &= n\hat{s}_n^2 K_n''(0) + n\hat{s}_n^3 O_{\mathbb{P}}(1). \end{aligned} \quad (189)$$

Proof of main results in Lemma 40: Now we come to prove the main results in Lemma 40 using the auxiliary results proved above.

Proof of (142) This can be directly obtained by (177) that

$$\frac{r_n^2}{n} = \hat{s}_n^2 K_n''(0) + \hat{s}_n^3 O_{\mathbb{P}}(1) = o_{\mathbb{P}}(1)$$

since $\hat{s}_n = o_{\mathbb{P}}(1)$ and by condition (138) and Cauchy-Schwarz inequality,

$$K_n''(0) = \frac{1}{n}\sum_{i=1}^n K_{in}''(0) \leq \left(\frac{1}{n}\sum_{i=1}^n (K_{in}''(0))^2\right)^{1/2} = O_{\mathbb{P}}(1).$$

Proof of (143) Since $\lambda_n, r_n \in (-\infty, \infty)$, we need to divide the proof into several cases. When $\hat{s}_n = 0$, we know $\lambda_n = r_n = 0$ so that $\lambda_n/r_n = 1$ by convention. Now we consider when $\hat{s}_n \neq 0$.

- **When** $\lambda_n r_n \leq 0$**:** observe that

$$\mathbb{P}\left[\frac{\mathbb{1}(\hat{s}_n \neq 0 \text{ and } \lambda_n r_n \leq 0)}{|\hat{s}_n|}\left|\frac{\lambda_n}{r_n} - 1\right| > \delta\right] \leq \mathbb{P}[\lambda_n r_n \leq 0 \text{ and } \hat{s}_n \neq 0] \to 0$$

so that

$$\mathbb{1}(\hat{s}_n \neq 0 \text{ and } \lambda_n r_n \leq 0) \left| \frac{\lambda_n}{r_n} - 1 \right| = \hat{s}_n O_{\mathbb{P}}(1).$$

- **When $\lambda_n r_n > 0$:** It requires to compute λ_n^2 / r_n^2. By (188) and (189), we get

$$\begin{aligned}
\frac{\lambda_n^2}{r_n^2} &= \frac{\hat{s}_n^2 K_n''(0) + \hat{s}_n^3 K_n'''(0) + \frac{\hat{s}_n^4}{2} K_n''''(\dot{s}(\hat{s}_n))}{\hat{s}_n^2 K_n''(0) + \hat{s}_n^3 M_n} \\
&= 1 + \hat{s}_n \frac{K_n'''(0) - \hat{s}_n M_n + \frac{\hat{s}_n}{2} K_n''''(\dot{s}(\hat{s}_n))}{K_n''(0) + \hat{s}_n M_n} \\
&\equiv 1 + \hat{s}_n \cdot F_n.
\end{aligned}$$

Thus we know

$$\mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) \left(\frac{\lambda_n^2}{r_n^2} - 1 \right) = \mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) \hat{s}_n \cdot F_n.$$

To further proceed the proof, we observe

$$F_n = O_{\mathbb{P}}(1)$$

since $\hat{s}_n = o_{\mathbb{P}}(1), M_n = O_{\mathbb{P}}(1)$ and conditions (141), (139) and (140) guarantee respectively $K_n''(0) = \Omega_{\mathbb{P}}(1), K_n'''(0) = O_{\mathbb{P}}(1)$ and $K_n''''(\dot{s}(\hat{s}_n)) = O_{\mathbb{P}}(1)$. Thus

$$\begin{aligned}
\mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) \left| \frac{\lambda_n}{r_n} - 1 \right| &\leq \mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) \left| \frac{\lambda_n}{r_n} - 1 \right| \left| \frac{\lambda_n}{r_n} + 1 \right| \\
&= \mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) \left| \frac{\lambda_n^2}{r_n^2} - 1 \right| \\
&\leq \mathbb{1}(r_n \lambda_n > 0 \text{ and } \hat{s}_n \neq 0) |\hat{s}_n F_n| \\
&= \hat{s}_n O_{\mathbb{P}}(1)
\end{aligned}$$

Collecting all the results, we have

$$\left| \frac{\lambda_n}{r_n} - 1 \right| = \hat{s}_n O_{\mathbb{P}}(1).$$

Proof of (144) The proof is similar to (143) so we omit the proof.

Proof of (145) When $\hat{s}_n = 0$, we know $\lambda_n = r_n = 0$ so that $1/\lambda_n - 1/r_n = 0$ by the convetion $1/0 - 1/0 = 1$. Now we consider the case when $\hat{s}_n \neq 0$. We divide the proof into serveral cases. By result (143) and claim (189),

$$\left(\frac{1}{r_n} - \frac{1}{\lambda_n} \right)^2 = \frac{1}{\lambda_n^2} \left(\frac{\lambda_n}{r_n} - 1 \right)^2 = \frac{\hat{s}_n^2 O_{\mathbb{P}}(1)}{n \hat{s}_n^2 K_n''(0) + n \hat{s}_n^3 O_{\mathbb{P}}(1)} = \frac{O_{\mathbb{P}}(1)}{n K_n''(0) + n \hat{s}_n O_{\mathbb{P}}(1)} = o_{\mathbb{P}}(1)$$

where the last equality is due to $\hat{s}_n = o_{\mathbb{P}}(1)$ and $K_n''(0) = \Omega_{\mathbb{P}}(1)$.

Proof of (146) On the event $r_n > 0, \lambda_n > 0$, we know $\hat{s}_n > 0$. Then by (143) and (188) we can compute

$$\begin{aligned}\mathbb{1}(r_n > 0 \text{ and } \lambda_n > 0)\frac{1}{r_n^2}\left(\frac{\lambda_n}{r_n} - 1\right)^2 &= \frac{\mathbb{1}(r_n > 0 \text{ and } \lambda_n > 0)\hat{s}_n^2 O_{\mathbb{P}}(1)}{n\hat{s}_n^2 K_n''(0) + n\hat{s}_n^3 O_{\mathbb{P}}(1)} \\ &= \frac{\mathbb{1}(r_n > 0 \text{ and } \lambda_n > 0)O_{\mathbb{P}}(1)}{nK_n''(0) + n\hat{s}_n O_{\mathbb{P}}(1)}.\end{aligned}$$

Then since $\hat{s}_n = o_{\mathbb{P}}(1)$ and $K_n''(0) = \Omega_{\mathbb{P}}(1)$, we know

$$\mathbb{1}(r_n > 0 \text{ and } \lambda_n > 0)\frac{1}{r_n}\left(\frac{\lambda_n}{r_n} - 1\right) = o_{\mathbb{P}}(1).$$

Proof of (147) The proof is similar to the proof of (146) so we omit it.

Proof of (148) Since $\mathbb{P}[\hat{s}_n > 0 \text{ and } \lambda_n r_n \leq 0] \to 0$ by (176) and $\operatorname{sgn}(w_n) = \operatorname{sgn}(\hat{s}_n)$, we have

$$\mathbb{P}[w_n > 0 \text{ and } \lambda_n r_n \leq 0] = \mathbb{P}[w_n > 0 \text{ and } \hat{s}_n > 0 \text{ and } \lambda_n r_n \leq 0] \to 0$$

Proof of (149) Since $\mathbb{P}[\hat{s}_n \neq 0 \text{ and } \lambda_n r_n \leq 0] \to 0$ by (176), we have

$$\mathbb{P}[\hat{s}_n \neq 0 \text{ and } \lambda_n r_n = 0] \leq \mathbb{P}[\hat{s}_n \neq 0 \text{ and } \lambda_n r_n \leq 0] \to 0.$$

□

I.20 Proof of Lemma 41

Proof of Lemma 41. We prove the statements (152)-(153) in order.

Proof of (152)**:** It suffices to prove that for any $\kappa > 0$,

$$\begin{aligned}&\mathbb{P}\left[\mathbb{1}(w_n < 0)\frac{\Phi(r_n) + \phi(r_n)\left\{\frac{1}{r_n} - \frac{1}{\lambda_n}\right\}}{1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}} \in [0, 1 + \kappa)\right] \\ &\equiv \mathbb{P}\left[\mathbb{1}(w_n < 0)A(r_n, \lambda_n) \in [0, 1 + \kappa)\right] \\ &\to 1. \end{aligned} \tag{190}$$

We decompose

$$\mathbb{1}(w_n < 0)A(r_n, \lambda_n) = \mathbb{1}(w_n < 0, r_n > 0)A(r_n, \lambda_n) + \mathbb{1}(w_n < 0, r_n \leq 0)A(r_n, \lambda_n).$$

By condition (Sign1), we know $\mathbb{1}(w_n < 0, r_n > 0) = 0$. Moreover, by the statement of (132) in Lemma 33, we know

$$\mathbb{P}\left[1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\} = 0\right] \to 0.$$

Therefore, for any $\delta > 0$,

$$\mathbb{P}\left[|\mathbb{1}(w_n < 0, r_n > 0)A(r_n, \lambda_n)| > \delta\right] \leq \mathbb{P}\left[1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\} = 0\right] \to 0.$$

Thus we know $\mathbb{1}(w_n < 0, r_n > 0)A(r_n, \lambda_n) = o_{\mathbb{P}}(1)$. Then we only need to consider behavior of $\mathbb{1}(w_n < 0, r_n \leq 0)A(r_n, \lambda_n)$. Then we know $1 - \Phi(r_n) \geq 1/2$ when $r_n \leq 0$. Then by condition (Rate1), we know

$$|M_n| \equiv \left|\phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}\right| \leq \left|\frac{1}{\lambda_n} - \frac{1}{r_n}\right| = o_{\mathbb{P}}(1).$$

Fix $\eta > 0, \delta \in (0, 0.1)$. Then for large enough $n, \mathbb{P}[|M_n| < \delta] \geq 1 - \eta$. Then on the event $|M_n| < \delta$ and $r_n \leq 0$, we have

$$\begin{aligned} 0 < \frac{\delta}{1-\delta} = \frac{1}{1-\delta} - 1 < A(r_n, \lambda_n) &= \frac{1}{1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}} - 1 \\ &\leq \frac{1}{\frac{1}{2} - \delta} - 1 = 1 + \frac{2\delta}{1 - 2\delta} < 1 + 4\delta. \end{aligned} \tag{191}$$

Thus we know

$$\liminf_{n\to\infty} \mathbb{P}\left[\mathbb{1}(w_n < 0, r_n \leq 0)A(r_n, \lambda_n) \in [0, 1 + 4\delta)\right] \geq \liminf_{n\to\infty} \mathbb{P}[|M_n| < \delta] > 1 - \eta.$$

Then by the arbitrary choice of η, we have

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}\left[\mathbb{1}(w_n < 0)A(r_n, \lambda_n) \in [0, 1 + 4\delta)\right] \\ &= \lim_{n\to\infty} \mathbb{P}\left[\mathbb{1}(w_n < 0 \text{ and } r_n \leq 0)A(r_n, \lambda_n) \in [0, 1 + 4\delta)\right] = 1. \end{aligned} \tag{192}$$

Thus we complete the proof for claim (190) by choosing $\kappa = 4\delta$.

Proof of (153): We have

$$\begin{aligned} &\mathbb{1}(w_n < 0)\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right]}{1 - \Phi(r_n) + \phi(r_n)\{\frac{1}{\lambda_n} - \frac{1}{r_n}\}} \\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right] \cdot \frac{\mathbb{1}(w_n < 0)}{1 - \Phi(r_n) + \phi(r_n)\left\{\frac{1}{\lambda_n} - \frac{1}{r_n}\right\}} \\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right] \cdot (\mathbb{1}(w_n < 0)A(r_n, \lambda_n) + \mathbb{1}(w_n < 0)) \\ &= \mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right] \cdot O_{\mathbb{P}}(1) \end{aligned}$$

where the second equality is due to the decomposition of $A(r_n, \lambda_n)$ in (191) and the last equality is due to result (192) that $\mathbb{1}(w_n < 0)A(r_n, \lambda_n) = O_{\mathbb{P}}(1)$. Now applying claim (117) with $y_n = w_n\sqrt{n/K_n''(0)}$, we know

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right] = o_{\mathbb{P}}(1).$$

Therefore we conclude

$$\mathbb{1}(w_n < 0)\frac{\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^n W_{in} = w_n \mid \mathcal{F}_n\right]}{1-\Phi(r_n)+\phi(r_n)\{\frac{1}{\lambda_n}-\frac{1}{r_n}\}} = o_{\mathbb{P}}(1).$$

□

I.21 Proof of Lemma 42

Proof of Lemma 42. We apply integration by parts to the following integral on the event $\lambda_n, r_n \in (-\infty, \infty)$,

$$\begin{aligned}
&\int_0^\infty \exp(-\lambda_n y)\phi(y)\mathrm{d}y \\
&= \exp(\lambda_n^2/2)(1-\Phi(\lambda_n)) \\
&= \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) + \int_{r_n}^{\lambda_n} y\exp(y^2/2)(1-\Phi(y)) - \frac{1}{\sqrt{2\pi}}\mathrm{d}y \\
&= \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) - \int_{r_n}^{\lambda_n} \frac{1}{\sqrt{2\pi}y^2}\mathrm{d}y + R_n \\
&= \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) + \frac{1}{\sqrt{2\pi}}\left(\frac{1}{\lambda_n}-\frac{1}{r_n}\right) + R_n.
\end{aligned}$$

This completes the proof. □

I.22 Proof of Lemma 43

Proof of Lemma 43. We prove the two claims separately.

Proof of (162): We can write (162) as

$$\frac{\mathbb{1}(r_n \geq 1)}{r_n U_n} = \frac{\mathbb{1}(r_n \geq 1)}{\mathbb{1}(r_n \geq 1) r_n \exp(\frac{1}{2}r_n^2)(1-\Phi(r_n)) + \frac{1}{\sqrt{2\pi}}\{\frac{r_n}{\lambda_n}-1\}}.$$

Notice by (Rate2),

$$\frac{1}{\sqrt{2\pi}}\left|\frac{r_n}{\lambda_n}-1\right| = o_{\mathbb{P}}(1).$$

Then it suffices to prove there exists a universal constant $C > 0$ such that

$$\mathbb{1}(r_n \geq 1) r_n \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) \geq C\mathbb{1}(r_n \geq 1).$$

To prove this, we apply Lemma 18 such that

$$r_n \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) \geq \frac{1}{\sqrt{2\pi}}\frac{r_n^2}{r_n^2+1} \geq \frac{1}{2}, \quad \text{when } r_n \geq 1.$$

Thus we have

$$\mathbb{1}(r_n \geq 1) r_n \exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) \geq \frac{\mathbb{1}(r_n \geq 1)}{2}.$$

Therefore we have proved (162).

Proof of (163): Similarly, we can write (163) as

$$\mathbb{1}(r_n \in [0,1))\frac{1}{U_n} = \frac{\mathbb{1}(r_n \in [0,1))}{\mathbb{1}(r_n \in [0,1))\exp(\frac{1}{2}r_n^2)(1-\Phi(r_n)) + \frac{1}{\sqrt{2\pi}}\{\frac{1}{\lambda_n} - \frac{1}{r_n}\}}.$$

By (Rate1),

$$\frac{1}{\sqrt{2\pi}}\left|\frac{1}{\lambda_n} - \frac{1}{r_n}\right| = o_{\mathbb{P}}(1).$$

We only need to prove there exists a universal constant $C \geq 0$ such that

$$\mathbb{1}(r_n \in [0,1))\exp\left(\frac{r_n^2}{2}\right)(1-\Phi(r_n)) \geq C\mathbb{1}(r_n \in [0,1)).$$

Indeed, we can set C to be

$$\inf_{z\in[0,1]} \exp\left(\frac{z^2}{2}\right)(1-\Phi(z)).$$

Therefore we proved claim (163). □

I.23 Proof of Lemma 44

Proof of Lemma 44. Using Lemma 42, we obtain for any $\lambda_n, r_n > 0$,

$$\int_0^\infty \exp(-\lambda_n y)\phi(y)\mathrm{d}y = R_n + U_n \tag{193}$$

almost surely. By statement (132) of Lemma 33, we know $\mathbb{1}(r_n > 0)/U_n \in (-\infty, \infty)$ almost surely and thus $\mathbb{1}(r_n > 0, \lambda_n > 0)/U_n \in (-\infty, \infty)$ almost surely. Then multiplying both sides in (193) with $\mathbb{1}(r_n > 0, \lambda_n > 0)/U_n$, we obtain

$$\mathbb{1}(r_n > 0, \lambda_n > 0)\frac{\int_0^\infty \exp(-\lambda_n y)\phi(y)\mathrm{d}y}{U_n} = \mathbb{1}(r_n > 0, \lambda_n > 0)(1 + R_n/U_n)$$

almost surely. This implies

$$\mathbb{1}(r_n > 0, \lambda_n > 0)\left|\frac{\int_0^\infty \exp(-\lambda_n y)\phi(y)\mathrm{d}y}{U_n} - 1\right| = \mathbb{1}(r_n > 0, \lambda_n > 0)\left|\frac{1}{U_n}\right| \cdot |R_n|$$

almost surely. Thus it suffcies to bound $\mathbb{1}(r_n > 0, \lambda_n > 0)|R_n/U_n|$. Define

$$R_{\min} \equiv \min\{r_n, \lambda_n\},\ R_{\max} \equiv \max\{r_n, \lambda_n\}.$$

Therefore the absolute value of R_n can be bounded as, using Lemma 17,

$$\begin{aligned}
&\mathbb{1}(r_n > 0, \lambda_n > 0)|R_n| \\
&= \mathbb{1}(\lambda_n > 0, r_n > 0)|R_n| \\
&\leq \mathbb{1}(\lambda_n > 0, r_n > 0)|r_n - \lambda_n| \sup_{y\in[R_{\min}, R_{\max}]}\left|y\exp(y^2/2)(1-\Phi(y)) - \frac{1-y^{-2}}{\sqrt{2\pi}}\right| \\
&\leq \mathbb{1}(\lambda_n > 0, r_n > 0)|r_n - \lambda_n|\frac{1}{\sqrt{2\pi}} \sup_{y\in[R_{\min}, R_{\max}]}\frac{1}{y^2} \\
&\leq \frac{\mathbb{1}(\lambda_n > 0, r_n > 0)|r_n - \lambda_n|}{\sqrt{2\pi}}\left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2}\right).
\end{aligned}$$

Then we have

$$\mathbb{1}(r_n > 0, \lambda_n > 0)|R_n| \leq \frac{\mathbb{1}(r_n > 0, \lambda_n > 0)|r_n - \lambda_n|}{\sqrt{2\pi}} \left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2} \right).$$

Then this implies

$$\mathbb{1}(r_n > 0, \lambda_n > 0) \left| \frac{R_n}{U_n} \right| \leq \left| \frac{1}{U_n} \right| \cdot \frac{\mathbb{1}(r_n > 0, \lambda_n > 0)|r_n - \lambda_n|}{\sqrt{2\pi}} \left(\frac{1}{r_n^2} + \frac{1}{\lambda_n^2} \right)$$

almost surely. Therefore we complete the proof. □

I.24 Proof of Lemma 45

Proof of Lemma 45. Then by Lemma 19 and Lemma 39, we have,

$$\mathbb{P}[\mathcal{T}] = 1, \ \mathcal{T} \equiv \left\{ K'_{in}(s) = \frac{\mathbb{E}[W_{in} \exp(sW_{in})|\mathcal{F}_n]}{\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n]}, \ \forall s \in (-\varepsilon, \varepsilon) \right\}.$$

Then we know,

$$K'_{in}(s)(\omega) = \frac{\mathbb{E}[W_{in} \exp(sW_{in})|\mathcal{F}_n]}{\mathbb{E}[\exp(sW_{in})|\mathcal{F}_n]}(\omega) = \mathbb{E}_{n,s}[W_{in}|\mathcal{F}_n](\omega), \ \forall \omega \in \mathcal{T}.$$

so that $\mathbb{P}[K'_{in}(s) = \mathbb{E}_{n,s}[W_{in}|\mathcal{F}_n], \ \forall s \in (-\varepsilon, \varepsilon)] = 1$. The other two claims follow similarly. □

I.25 Proof of Lemma 46

Proof of Lemma 46. We use the definition of conditional expectation (32) to prove the claim:

$$\begin{aligned}
\sum_{m=0}^{\infty} \frac{\mathbb{E}[|W_{in}|^m|\mathcal{F}_n]}{m!}|s|^m &= \sum_{m=0}^{\infty} \frac{s^m}{m!} \int x^m \mathrm{d}\kappa_{in}(\cdot, x) \\
&= \int \sum_{m=0}^{\infty} \frac{s^m}{m!} x^m \mathrm{d}\kappa_{in}(\cdot, x) \\
&= \int \exp(sx) \mathrm{d}\kappa_{in}(\cdot, x)
\end{aligned}$$

where the second equality is due to Fubini's theorem. Thus we have proved the claim. Then for the inequality $\mathbb{E}[\exp(|sW_{in}|)|\mathcal{F}_n] < \infty$, we can bound, on the event $\mathcal{A}$,

$$\mathbb{E}[\exp(|sW_{in}|)|\mathcal{F}_n] \leq \mathbb{E}[\exp(sW_{in})|\mathcal{F}_n] + \mathbb{E}[\exp(-sW_{in})|\mathcal{F}_n] < \infty, \forall s \in (-\varepsilon, \varepsilon).$$

Thus we have proved the claim. □

J Proof of Theorem 5

The proof follows by applying Theorem 1 with

$$W_{in} = \widetilde{X}_{in},\ w_n = \frac{1}{n}\sum_{i=1}^{n} X_{in},\ \mathcal{F}_n = \mathcal{G}_n,\ \varepsilon = 2.$$

Thus it suffices to verify the assumptions required in Theorem 1 with these realizations. In particular, we will verify $W_{in}|\mathcal{F}_n$ in this case satisfies **CCS condition**. Notice

$$\mathbb{P}[W_{in} \in [-|X_{in}|, |X_{in}|]|\mathcal{F}_n] = 1,\ \text{almost surely.}$$

Then by Theorem 1, it suffices to verify the following lemma:

Lemma 47. *Suppose the assumptions of Theorem 5 hold. Then*

$$\frac{1}{n}\mathbb{E}[W_{in}^2|\mathcal{F}_n] = \frac{1}{n}\sum X_{in}^2 = \Omega_{\mathbb{P}}(1),\ \frac{1}{n}\sum_{i=1}^{n} X_{in}^4 = O_{\mathbb{P}}(1).$$

We now conclude this section by proving Lemma 47.

Proof of Lemma 47. We will apply Lemma 11 to prove the claims.

Proof of $\sum_{i=1}^n X_{in}^4/n = O_{\mathbb{P}}(1)$: We will apply Lemma 11 with $W_{in} = X_{in}^4$ and $\kappa = \delta/4$. If we can verify,

$$\frac{1}{n^{1+\delta/4}}\sum_{i=1}^{n}\mathbb{E}[|X_{in}^4|^{1+\delta/4}] = \frac{1}{n^{1+\delta/4}}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^{4+\delta}] \to 0.$$

then applying Lemma 11, we have

$$\frac{1}{n}\sum_{i=1}^{n}(X_{in}^4 - \mathbb{E}[X_{in}^4]) = o_{\mathbb{P}}(1).$$

It suffices to show

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_{in}^4] < \infty,\ \limsup_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^{4+\delta}] < \infty. \tag{194}$$

By Lemma (15), it suffices to just show the $4+\delta$ moment condition in (54). In fact, using the inequality $(|a|+|b|)^p \le 2^p(|a|^p + |b|^p)$ for $p > 0$, we can bound

$$|X_{in}|^{4+\delta} = |\mu_n + \varepsilon_{in}|^{4+\delta} \le 2^{4+\delta}|\mu_n|^{4+\delta} + 2^{4+\delta}|\varepsilon_{in}|^{4+\delta}$$

so that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[|X_{in}|^{4+\delta}] \le 2^{4+\delta}|\mu_n|^{4+\delta} + \frac{2^{4+\delta}}{n}\sum_{i=1}^{n}\mathbb{E}[|\varepsilon_{in}|^{4+\delta}].$$

By condition (55) and condition (54), we know the claim is true. Therefore, we have

$$\frac{1}{n}\sum_{i=1}^{n}X_{in}^4 = \frac{1}{n}\sum_{i=1}^{n}(X_{in}^4 - \mathbb{E}[X_{in}^4]) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_{in}^4] = O_{\mathbb{P}}(1).$$

Proof of $\sum_{i=1}^n X_{in}^2/n = \Omega_{\mathbb{P}}(1)$: We will apply Lemma 11 with $W_{in} = X_{in}^2$ and $\kappa = 1$. In other words, we need to verify

$$\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[X_{in}^4] \to 0.$$

This is true by cliam (194) that $\limsup_{n\to\infty} \sum_{i=1}^n \mathbb{E}[X_{in}^4]/n < \infty$. Therefore applying Lemma 11, we have

$$\frac{1}{n}\sum_{i=1}^n (X_{in}^2 - \mathbb{E}[X_{in}^2]) = o_{\mathbb{P}}(1).$$

Thus in order to prove $\sum_{i=1}^n X_{in}^2/n = \Omega_{\mathbb{P}}(1)$, it suffices to show $\liminf_{n\to\infty} \sum_{i=1}^n \mathbb{E}[X_{in}^2]/n > 0$. Indeed,

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_{in}^2] = \mu_n^2 + \frac{2\mu_n}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_{in}] + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_{in}^2] \geq \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_{in}^2].$$

where the second inequality is true because $\mathbb{E}[\varepsilon_{in}] = 0$ due to the symmetric distribution assumption. By condition (53), we know

$$\liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_{in}^2] \geq \liminf_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\varepsilon_{in}^2] > 0.$$

Therefore

$$\frac{1}{n}\sum_{i=1}^n X_{in}^2 = \frac{1}{n}\sum_{i=1}^n (X_{in}^2 - \mathbb{E}[X_{in}^2]) + \frac{1}{n}\sum_{i=1}^n \mathbb{E}[X_{in}^2] = \Omega_{\mathbb{P}}(1).$$

□

K Proof of Theorem 2 and Corollary 2

K.1 Proof of Theorem 2

We have conditional CGF

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

Then we can compute the CCGF under model (13) as

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

The first two derivatives of this quantity are

$$K_n'(s \mid \mathcal{F}_n) = \frac{1}{n}\sum_{i=1}^n a_{in}\left(A'(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s) - A'(\widehat{\theta}_{n,x}(Z_{in})) \right), \tag{195}$$

$$K_n''(s \mid \mathcal{F}_n) = \frac{1}{n}\sum_{i=1}^n a_{in}^2 A''(\widehat{\theta}_{n,x}(Z_{in}) + a_{in}s). \tag{196}$$

We will apply Theorem 1 and thus verify the conditions in the theorem. We first verify the variance condition (7).

Verification of (7): 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 Assumption 2 in Theorem 1 with condition (CSE) and (CCS), respectively.

Verification of Assumption 1 with condition (CSE) and (19): We denote the conditional upper tail probability and lower probability respectively as

$$L_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}\left[X \le a|\mathcal{F}_n\right] \quad \text{and} \quad U_{X,\mathcal{F}_n}(a) \equiv \mathbb{P}\left[X \ge a|\mathcal{F}_n\right].$$

By condition (CSE), we can compute

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

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

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

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

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

Then we have for any $t > 0$,

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

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 (19), we know $\sup_i |a_{in}| \leq \sup_i |Y_{in}| + \sup_i |\widehat{\mu}_{n,y}(Z_{in})| < \infty$ 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

$$
\mathbb{P}[\mathcal{S}] \geq 1 - \delta, \text{ where } \mathcal{S} \equiv \left\{\sup_i |\widehat{\theta}_{n,x}(Z_{in})|, \sup_i |a_{in}| \in [0, M(\delta)]\right\}.
$$

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

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

and

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

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

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

Therefore we have

$$
\mathbb{P}\left[\theta_n \leq \exp\left(2 \sup_{x \in [-2M(\delta), 2M(\delta)]} |A(x)| + M(\delta) \sup_{x \in [-M(\delta), M(\delta)]} |A'(x)|\right)\right] \geq \mathbb{P}[\mathcal{S}] \geq 1 - \delta.
$$

Therefore we have $\theta_n = O_{\mathbb{P}}(1)$. Thus ε in Lemma 5 can be chosen to be $\beta/8 = 1/8$, according to the proof of Lemma 5.

Verification of Assumption 2 with condition (19) 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

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

Then since $\widehat{\mu}_{n,x}(Z_{in}) = \mathbb{E}[\widetilde{X}_{in} | \mathcal{F}_n]$, we have $W_{in} = a_{in}(\widetilde{X}_{in} - \widehat{\mu}_{n,x}(Z_{in}))$ so that

$$\mathbb{P}\left[W_{in} \in [-2|a_{in}|S, 2|a_{in}|S] | \mathcal{F}_n\right] = 1, \text{ almost surely.}$$

Then again by condition (19), we know

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

Moreover, by condition (CCS), we know

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

Choosing $\nu_{in} = 2|a_{in}|S$, we complete the proof for CCS distribution. Thus ε can be chosen to be $1/8$ according to the proof of Lemma 5.

K.2 Proof of Corollary 2

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

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

By the asymptotic validity of dCRT, $\lim_{n\to\infty} \mathbb{P}_{H_0}[p_{\text{dCRT}} \leq \alpha] \leq \alpha$, and conclusion (22) in Theorem 2, we have

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

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

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

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

L Proof of Theorem 3

We divide the proof into two parts. First, we prove that the conclusion of Theorem 2 is correct. Then, we show that spaCRT controls Type-I error asymptotically. Since we assume fixed-dimensional setup, we drop the subscript n for notational simplicity.

L.1 Proof of the conclusion in Theorem 2

We need to apply Lemma 7 with **Condition set 2**. In particular, the condition (74) is trivially satisfied by the finiteness of the maximum likelihood estimate $\widehat{\beta}, \widehat{\gamma}$. Also condition (75) is true in the low-dimensional setup. Thus it suffices to prove condition (71). We divide the proof into two steps.

1. **Verificaiton of $\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i) - \widehat{\mu}_y(Z_{in}))^4 = o_{\mathbb{P}}(1)$.** We will prove a stronger result

$$\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^4 = O_{\mathbb{P}}(1/n^2). \tag{197}$$

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

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

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

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}(\mu_y(Z_i) - \widehat{\mu}_y(Z_i))^4 &= \frac{1}{n}\sum_{i=1}^{n}(A'(Z_i^\top\beta) - A'(Z_i^\top\widehat{\beta}))^4 \\
&\le \frac{1}{n}\sum_{i=1}^{n}\sup_{t\in C}(A''(t))^4 |Z_i^\top(\beta - \widehat{\beta})|^4 \\
&\qquad \text{(Mean value theorem)} \\
&\le \sup_{t\in C}(A''(t))^4 \sup_i \|Z_i\|_\infty^4 \|\widehat{\beta} - \beta\|_1^4 \\
&\qquad \text{(Hölder's inequality)} \\
&\le \sup_{t\in C}(A''(t))^4 C_Z^4 \|\widehat{\beta} - \beta\|_1^4 \qquad \text{(Assumption 4)} \\
&= O_{\mathbb{P}}(1/n^2). \qquad \text{(Condition (23))}
\end{aligned}$$

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

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

By condition (23), we have $\|\widehat{\beta} - \beta\|_1 = o_{\mathbb{P}}(1)$ and $\|\widehat{\gamma} - \gamma\|_1 = o_{\mathbb{P}}(1)$. Thus $\mathbb{P}[\mathcal{C}] \to 1$.

2. **Verification of $\frac{1}{n}\sum_{i=1}^n(\theta(Z_i)-\widehat{\theta}_x(Z_i))^2=o_{\mathbb{P}}(1)$.** We will also prove a stronger result:

$$\frac{1}{n}\sum_{i=1}^n(\theta(Z_i)-\widehat{\theta}_x(Z_i))^2=\frac{1}{n}\sum_{i=1}^n(Z_i^\top\widehat{\gamma}-Z_i^\top\gamma)^2=O_{\mathbb{P}}(1/n).$$

By Hölder's inequality, we have

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

L.2 Proof of the asymptotic validity under null

We have verified the conditions for Lemma 7 to hold in section L.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 Theorem 8, it is sufficient to prove

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

We will apply Theorem 16 to prove the claim. 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 second claim, by Cauchy-Schwarz inequality and result (197), we have

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

For the first claim, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n(\mu_x(Z_i)-\widehat{\mu}_x(Z_i))^2&=\frac{1}{n}\sum_{i=1}^n(\mathrm{expit}(Z_i^\top\gamma)-\mathrm{expit}(Z_i^\top\widehat{\gamma}))^2\\&\le\frac{1}{n}\sum_{i=1}^n(Z_i^\top\gamma-Z_i^\top\widehat{\gamma})^2\quad\text{(By Lipschitz continuity of expit)}\\&\le C_Z^2\|\widehat{\gamma}-\gamma\|_1^2=O_{\mathbb{P}}(1/n).\quad\text{(By Assumption 4 and (23))}\end{aligned}$$

Therefore we complete the proof.

M Proof of Theorem 4

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

$$\widehat{\beta}_n=\arg\min_\beta\left\{\frac{1}{n}\sum_{i=1}^n\{A_y(Z_{in}^\top\beta)-Y_{in}\cdot(Z_{in}^\top\beta)\}+\lambda_n\|\beta\|_1\right\}\tag{198}$$

and

$$\widehat{\gamma}_n=\arg\min_\gamma\left\{\frac{1}{n}\sum_{i=1}^n\{\log(1+\exp(Z_{in}^\top\gamma))-X_{in}\cdot(Z_{in}^\top\gamma)\}+\nu_n\|\gamma\|_1\right\}.\tag{199}$$

M.1 Proof of the conclusion in Theorem 2

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

Lemma 48. *Suppose all the assumptions in Theorem 4 except for (28) hold. Recall λ_n and ν_n as in models (198) and (199). 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.

The proof of Lemma 48 will be postponed to section M.3.

M.2 Proof of the asymptotic validity under null

By Lemma 48 and Theorem 8, in order to prove the validity of spaCRT, it suffices to show the following conditions are satisfied:

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

In fact, the variance convergence $(\widehat{S}_n^{\mathrm{dCRT}})^2$ has been proved as in Lemma 48. By Assumption 3, we know it suffices to prove

$$\frac{n^{1/2} T_n^{\mathrm{dCRT}}(X, Y, Z)}{\sqrt{\mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^2 (X_{in} - \mu_{n,x}(Z_{in}))^2]}} \xrightarrow{d} N(0, 1) \tag{200}$$

Proof of weak convergence (200). We 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

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

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

Lemma 49 (Bias term convergence)**.** *Suppose all the assumptions in Theorem 4 hold. Then we have* $\mathrm{Bias}_1, \mathrm{Bias}_2, \mathrm{Bias}_3 = o_{\mathbb{P}}(1)$.

Therefore, by Assumption 3, we just need to prove

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

Now we finish the proof by applying Lemma 26 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

$$\begin{aligned}\mathbb{E}[(X_{in} - \mu_{n,x}(Z_{in}))^4(Y_{in} - \mu_{n,y}(Z_{in}))^4] &\leq \mathbb{E}[(Y_{in} - \mu_{n,y}(Z_{in}))^4] \\ &= \mathbb{E}[A_y^{(4)}(Z^\top \gamma_n) + 3(A_y''(Z^\top \gamma_n))^2] \\ &\leq \sup_{\|t\|_2 \leq C_Z \sup_n \|\gamma_n\|_1} (A_y^{(4)}(t) + 3(A_y''(t))^2) < \infty\end{aligned}$$

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

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

Then we know (70) is satisfied. Thus by Theorem 8, we know

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

M.3 Proof of Lemma 48

We divide this section to two parts. We first introduce a strong consistency result for lasso estimator in Section M.3.1 and then prove the main result in Section M.3.2.

M.3.1 A strong consistency result for the lasso estimators

Proof of Lemma 48 hinges on a general consistency results proved in (Wainwright, 2019).

Lemma 50 (A modified version of Corollary 9.26 in (Wainwright, 2019))**.** *Consider the lasso estimators (198) and (199). Suppose assumptions 4-5 and condition (26) 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 whenever $n \geq N(\varepsilon)$, we have*

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

and

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

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

We now give the proof of the lemma.

Proof of Lemma 50. The proof requires Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019).

Lemma 51 (Corollary 9.26 and Theorem 9.36 in (Wainwright, 2019))**.** *Consider the lasso estimators (198) and (199). Suppose Assumptions 4-5 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 we have*

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

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 50, it is sufficient to show for some $\delta \in (0,1)$,

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

These conditions are satisfied by condition (26). □

M.3.2 Proof of Lemma 48

We prove the results by applying Lemma 7 with **Condition set 1**, combined with the result in Lemma 50. We need to verify conditions $T_n^{\mathrm{dCRT}}(X,Y,Z) \xrightarrow{\mathbb{P}} 0$, (71)-(73).

Verificaiton of $T_n^{\mathrm{dCRT}}(X,Y,Z) \xrightarrow{\mathbb{P}} 0$. 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 $B_n = o_{\mathbb{P}}(1)$.

Proof of $A_n = o_{\mathbb{P}}(1)$. Observe that A_n is just a sample average of conditionally independent random variables on data (X,Z). Define the set

$$B \equiv \{t \in \mathbb{R} : |t| \leq \sup_n \|\beta_n\|_1 C_Z + 1\}. \tag{201}$$

Then we have

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

then we know $\mathbb{E}[A_n^2] \le \max_{t\in B} A_y''(t)/n$. By conditional Markov's inequality (Lemma 14): for any $\varepsilon > 0$,

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

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

Proof of $B_n = o_{\mathbb{P}}(1)$. We observe

$$|\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in})| = |A_y'(Z_{in}^\top \beta_n) - A_y'(Z_{in}^\top \widehat{\beta}_n)|. \tag{202}$$

Recall that $B \equiv \{t \in \mathbb{R} : |t| \le \sup_n \|\beta_n\|_1 C_Z + 1\}$. Then, on the event

$$E_n \equiv \left\{ Z_{in}^\top \beta_n, Z_{in}^\top \widehat{\beta}_n \in B \text{ for any } i \in [n] \right\}, \tag{203}$$

we know

$$\begin{aligned}
|B_n| &\le \frac{1}{n}\sum_{i=1}^n |b_{in}(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && \text{(Triangle inequality)} \\
&\le \frac{1}{n}\sum_{i=1}^n |(\mathbb{E}[Y_{in}|Z_{in}] - \widehat{\mu}_{n,y}(Z_{in}))| && (|b_{in}| \le 1 \text{ almost surely}) \\
&= \frac{1}{n}\sum_{i=1}^n |A_y'(Z_{in}^\top \beta_n) - A_y'(Z_{in}^\top \widehat{\beta}_n)| && \text{(By result (202))} \\
&\le \frac{1}{n}\sum_{i=1}^n \max_{t\in B} A_y''(t) |Z_{in}^\top (\widehat{\beta}_n - \beta_n)| && \text{(Mean value theorem)} \\
&\le \max_{t\in B} A_y''(t) C_Z \|\widehat{\beta}_n - \beta_n\|_1. && \text{(Hölder's inequality)}
\end{aligned}$$

By Lemma 50, 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 52. *Suppose the assumptions in Theorem 48 hold. Then*

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

Proof of Lemma 52 can be found in section M.5. With Lemma 52, we conclude the verificaiton for (21).

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, on the event E_n (defined in (203)),

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

Thus we know on the event E_n,

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

Thus by Lemma 50, 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 since E_n^c will happen with probability 0 when n is large, guaranteed by Lemma 52. 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 50. 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,

$$\begin{aligned}|\widehat{\theta}_{n,x}(Z_{in})-\theta(Z_{in})| &= |\mathrm{expit}(Z_{in}^\top\gamma_n)-\mathrm{expit}(Z_{in}^\top\widehat{\gamma}_n)|\\ &\le |Z_{in}^\top\widehat{\gamma}_n-Z_{in}^\top\gamma_n|\le C_Z\|\widehat{\gamma}_n-\gamma_n\|_1\to 0\end{aligned}$$

almost surley by Lemma 50. 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:

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

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

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

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

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

This can be shown by noticing

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

and

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

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).

M.4 Proof of Lemma 49

For Bias_1, we first compute

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

Similarly, for Bias_2, we have

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

Then for any $\varepsilon > 0$, we have

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

and

$$\mathbb{P}[\text{Bias}_2 \ge \varepsilon] = \mathbb{P}[\text{Bias}_2 \wedge \varepsilon \ge \varepsilon] \le \frac{\mathbb{E}[\mathbb{E}[\text{Bias}_2^2|X,Z]\wedge\varepsilon]}{\varepsilon^2}.$$

Then by dominated convergence theorem, we know $\text{Bias}_1 = o_{\mathbb{P}}(1)$ and $\text{Bias} = o_{\mathbb{P}}(1)$ if

$$\frac{1}{n}\sum_{i=1}^n (\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2 = o_{\mathbb{P}}(1), \quad \frac{1}{n}\sum_{i=1}^n (\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^2 = o_{\mathbb{P}}(1). \tag{204}$$

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 we have $\text{Bias}_3 = o_{\mathbb{P}}(1)$ if

$$\left(\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,x}(Z_{in}) - \widehat{\mu}_{n,x}(Z_{in}))^2\right) \cdot \left(\frac{1}{n}\sum_{i=1}^{n}(\mu_{n,y}(Z_{in}) - \widehat{\mu}_{n,y}(Z_{in}))^2\right) = o_{\mathbb{P}}(1/n). \tag{205}$$

Now we prove claims (204) and (205). We compute

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

Then by Lemma 50, we know $\|\widehat{\gamma}_n - \gamma_n\|_1^2 = O_{\mathbb{P}}(s_{\gamma_n}^2 \log(d)/n)$. Thus we have

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

We can compute

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

where the last inequality is due to mean value theorem and B is defined as in (201). Therefore combining the above results and conditions (26) and (28), we know claims (204) and (205) hold so that we finish the proof.

M.5 Proof of Lemma 52

Proof of Lemma 52. First consider $Z_{in}^\top\beta_n$. By Hölder's iequality, we have

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

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 50, we know

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

Thus we know

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

By Lemma 50, we know there exists $N(C_Z) \in \mathbb{N}_+$ such that for any $n \ge N(C_Z)$,

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

Then we have

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

This concludes the proof. □

N Proof of results in Section E

N.1 Proof of Theorem 8

Proof of Theorem 8. Define the sequence M_n to be 0 if $p_{\text{spaCRT}} = 0$ and

$$M_n = \frac{\mathbb{P}\left(T_n^{\text{dCRT}}(\widetilde{X}, X, Y, Z) \ge T_n^{\text{dCRT}}(X, Y, Z) | \mathcal{F}_n\right)}{p_{\text{spaCRT}}} - 1 \quad \text{otherwise.}$$

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 (90). Then the remaining proof is divided as follows: (1) we prove the equivalence of $\phi_{n,\alpha}^{\text{aux}}$ and $\phi_{n,\alpha}^{\text{asy}}$, (2) we prove the equivalence of $\phi_{n,\alpha}^{\text{asy}}$ and $\phi_{n,\alpha}^{\text{dCRT}}$, (3) we prove the equivalence of $\phi_{n,\alpha}^{\text{aux}}$ and $\phi_{n,\alpha}^{\text{spaCRT}}$.

1. **Proof of the equivalence of $\phi_{n,\alpha}^{\text{aux}}, \phi_{n,\alpha}^{\text{asy}}$.** We apply Lemma 8 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

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

We will use Lemma 10 to prove the convergence of $C_n(X, Y, Z)$. To this end, we will first verify condition (87)-(89) in **Regularity condition** are satisfied under the assumptions of Theorem 8.

Verification of (87): This is true by assumption (20).

Verification of (88): We verify the condition when $\delta = 2$. We divide the proof to two cases: when condition (CSE) or (CCS) holds.

- **When condition (CSE) holds.** , it suffices to prove

$$\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}_{\widehat{\mathcal{L}}_n}[|\widetilde{X}_{in}-\widehat{\mu}_{n,x}(Z_{in})|^4|X,Z] = o_{\mathbb{P}}(1). \tag{206}$$

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 $\varepsilon > 0$ such that

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

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

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

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

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

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 by Lemma 2 we have

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

Therefore, we have proved (206) holds.

- **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 (89): $\mathrm{Var}_{\widehat{\mathcal{L}}_n}[X_{in}|Z_{in}] = A''(\widehat{\theta}(Z_{in})) < \infty$ 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 (19). As for $(Y_{in} - \mu_{n,y}(Z_{in}))^2 < \infty$, it is true by the integrability of Y_{in}.

Therefore, applying Lemma 10, 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 (86) holds for the chosen test statistic guaranteed by condition (70) so that this proves the asymptotic equivalence of $\phi_{n,\alpha}^{\mathrm{aux}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$.

2. **Proof of the equivalence of $\phi_{n,\alpha}^{\mathrm{asy}}, \phi_{n,\alpha}^{\mathrm{dCRT}}$.** In order to prove the asymptotic equivalence between $\phi_{n,\alpha}^{\mathrm{dCRT}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$, we apply Lemma 8 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 9, we have proved that under the assumptions in Theorem 8, $C_n(X, Y, Z) \xrightarrow{\mathbb{P}} z_{1-\alpha}$. Similarly, the nonaccumulant assumption (86) has been satisfied by (70) so that we have proved the asymptotic equivalence between $\phi_{n,\alpha}^{\mathrm{dCRT}}$ and $\phi_{n,\alpha}^{\mathrm{asy}}$.

3. **Proof of the equivalence of $\phi_{n,\alpha}^{\mathrm{aux}}, \phi_{n,\alpha}^{\mathrm{spaCRT}}$.** Notice the tests $\phi_{n,\alpha}^{\mathrm{spaCRT}}$ and $\phi_{n,\alpha}^{\mathrm{aux}}$ are equivalent as long as $M_n \in (-1, 1/\alpha - 1), p_{\mathrm{spaCRT}} \neq 0$ and $\widehat{S}_n^{\mathrm{dCRT}} \neq 0$. Indeed, by Theorem 2 and conclusion (87), we know $M_n = o_{\mathbb{P}}(1), \mathbb{P}[p_{\mathrm{spaCRT}} = 0] \to 0$ and $\mathbb{P}[\widehat{S}_n^{\mathrm{dCRT}} = 0] \to 0$, respectively. Therefore we have

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

This proves the asymptotic equivalence of $\phi_{n,\alpha}^{\mathrm{aux}}$ and $\phi_{n,\alpha}^{\mathrm{spaCRT}}$.

□

N.2 Proof of Lemma 7

We now divide the proof of Lemma 7 into three parts, depending either the **Condition set 1**, **Condition set 2**, or **Condition set 3** is used.

N.2.1 Proof of Lemma 7 with Condition set 1

Since $\boldsymbol{X}$ is a binary variable and follws model (13), 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 (19): 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

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

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

$$|Y_{in} - \mu_{n,y}(Z_{in})| < \infty, \forall i \in [n] \text{ almost surely.} \tag{207}$$

This is true by using Jensen's inequality (Lemma 13) and Cauchy-Schwarz inequality:

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

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

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

almost surely.

Verification of (20): We can write

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

We now divide the proof into two parts: $T_1 = \Omega_{\mathbb{P}}(1)$ and $T_2 = o_{\mathbb{P}}(1)$.

1. **Proof of $T_1 = \Omega_{\mathbb{P}}(1)$.** 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

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

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 11 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 11 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).

2. **Proof of** $T_2 = o_{\mathbb{P}}(1)$**.**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

$$\frac{1}{n}\sum_{i=1}^{n} a_{in}^4 = O_{\mathbb{P}}(1),\ \ \frac{1}{n}\sum_{i=1}^{n}|\widehat{\theta}_{n,x}(Z_{in}) - \theta(Z_{in})|^2 = o_{\mathbb{P}}(1). \tag{208}$$

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 13):

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

Therefore, we have proved $T_2 = o_{\mathbb{P}}(1)$.

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 (208).

N.2.2 Proof of Lemma 7 with Condition set 2

Checking the proof with **Condition set 1**, we know the proof for condition (19) is the only part that differs. However, condition (19) has been directly assumed in condition (74). Therefore we complete the proof.

N.2.3 Proof of Lemma 7 with Condition set 3

Verificaitons of conditions (19) and (CCS) are straightforward. As for condition (20), the proof can go through as that with **Condition set 1** by noting

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

Thus we complete the proof.

O 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 O.1 and section O.2 respectively. Before proceeding to the proof, we present a key lemma which states the consistency of the KRR estimator, provided in the proof of Theorem 11 in Shah and Peters (2020).

Proposition 3 (KRR consistency)**.** *Suppose the conditions in Theorem 9 hold. Then we have* $\frac{1}{n}\sum_{i=1}^n (\widehat{\mu}_y(Z_i) - \mu_y(Z_i))^2 = o_{\mathbb{P}}(1)$*, where* $\widehat{\mu}_y$ *is defined as in* (80)*.*

O.1 Proof of the conclusion in Theorem 2

We will verify the conditions in the version of Lemma 7 with **Condition set 3**. In fact, condition (77) is clearly satisfied by the boundedness of $\boldsymbol{Y}$ (condition (82)) and we now prove the other conditions in the lemma.

Verification of condition (71) We first show, by Corollary 3 and condition (84),

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

Then we show

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(\theta(Z_{in})-\widehat{\theta}_{n,x}(Z_{in}))^2 &= \frac{1}{n}\sum_{i=1}^{n}(Z_i^\top\widehat{\gamma}-Z_i^\top\gamma)^2 \\ &\le \sup_i \|Z_i\|_\infty^2\|\widehat{\gamma}-\gamma\|_1^2 && \text{(By Hölder's inequality)} \\ &\le C_Z^2\|\widehat{\gamma}-\gamma\|_1^2 && \text{(By Assumption 4)} \\ &= o_{\mathbb{P}}(1). && \text{(By condition (83))}\end{aligned}$$

This completes the verification of condition (71).

Verificaiton of condition (76) The $\widehat{\theta}_{n,x}(Z_{in}) = Z_i^\top\widehat{\gamma}$ 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 (21) We prove a stronger result:

$$\sqrt{n}T_n^{\mathrm{dCRT}} \xrightarrow{d} N(0,\sigma_{\mathrm{dCRT}}^2) \quad \text{where} \quad \sigma_{\mathrm{dCRT}}^2 = \mathbb{E}[(X_i-\mu_x(Z_i))^2(Y_i-\mu_y(Z_i))^2]. \quad (209)$$

By Lemma 16, it suffices to show the following

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

and

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

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

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

For the second claim, we know it is true by Proposition 3. Next, the third claim is obvious by the above arugments and condition (83).

O.2 Proof of the asymptotic validity under null

By Theorem 8, we just need to show that

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

We have proved the conditions required in Lemma 7 in section O.1 so we know by the conclusion (78) that

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

We have proved the conclusion (209) so by Slutsky's theorem, the desired claim is true.

P Additional simulation details in Section 5.1

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

In single-cell CRISPR screens, 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).

P.2 Parameters and methods implementation in Section 5.1

Parameters used in Section 5.1 We adopt the parameter settings displayed in Table 2. 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. When $\gamma_0 = -5$ and $\beta_0 = -5$, the marginal means of X and Y are approximately 0.01. When $\gamma_0 = -6$, the rate of X being nonzero is around 0.004. We comment that the choice of (γ_0, β_0) in our simulation study reflects empirically observed sparsity levels (see Figure 16 in Section R.2).

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 2: 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 3) 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 of spaCRT 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

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

  where $T_n^{\text{dCRT}}(X, Y, Z)$ is defined as in (14) 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 a more sophisticated estimator that requires iterative computation. Thus it is slower than the precomputed approach but more accurate. We choose this iterative approach since the score test relies more heavily on the accuracy of the size parameter estimate.

The comparison of different methods applied is summarized in Table 3. 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 3: Summary table for testing methods compared.

P.3 Additional simulation results in Section 5.1

The organization of this section is as follows.

1. **Multiplicity corrected results.** We present the rejection results when multiplicity is corrected using the Benjamini-Hochberg (BH) and Bonferroni methods in Figure 8.

2. **Size parameter varied results.** Next, we present the Type-I error control when varying the size parameter r in Figure 9-11. We also find that larger size parameters r lead to better behavior for the GCM and score tests. This is also to be expected, because smaller size parameters make the negative binomial distribution more skewed, and therefore the central limit theorem converges more slowly. Furthermore, smaller size parameters are more difficult to estimate accurately due to the increased variance in the gene expression Y, which impacts the score test. On the other hand, the dCRT and spaCRT behave much more stably across different sparsity levels of X and Y and different values of the size parameter r, since these methods do not rely on the central limit theorem.

3. **Approximation accuracy results.** To demonstrate the approximation accuracy of spaCRT to dCRT, we plot the the p-values comparison in Figure 12. We compare all the null p-values, stratified by the size parameter r. We find that p-values from spaCRT approximate dCRT very well for all values of r.

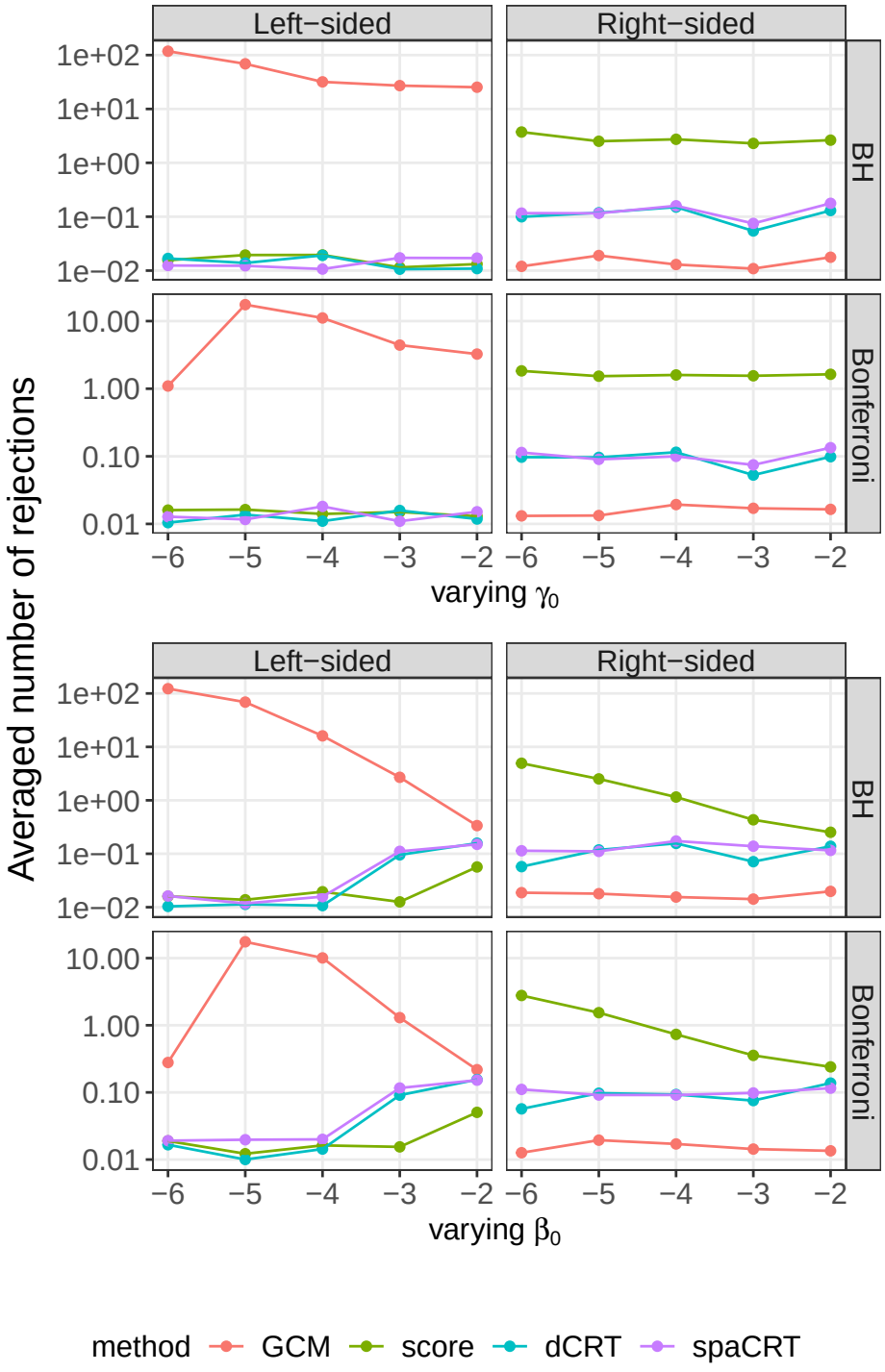

Figure 8: Averaged number of rejections after BH and Bonferroni corrections under the setup $r = 0.05$.

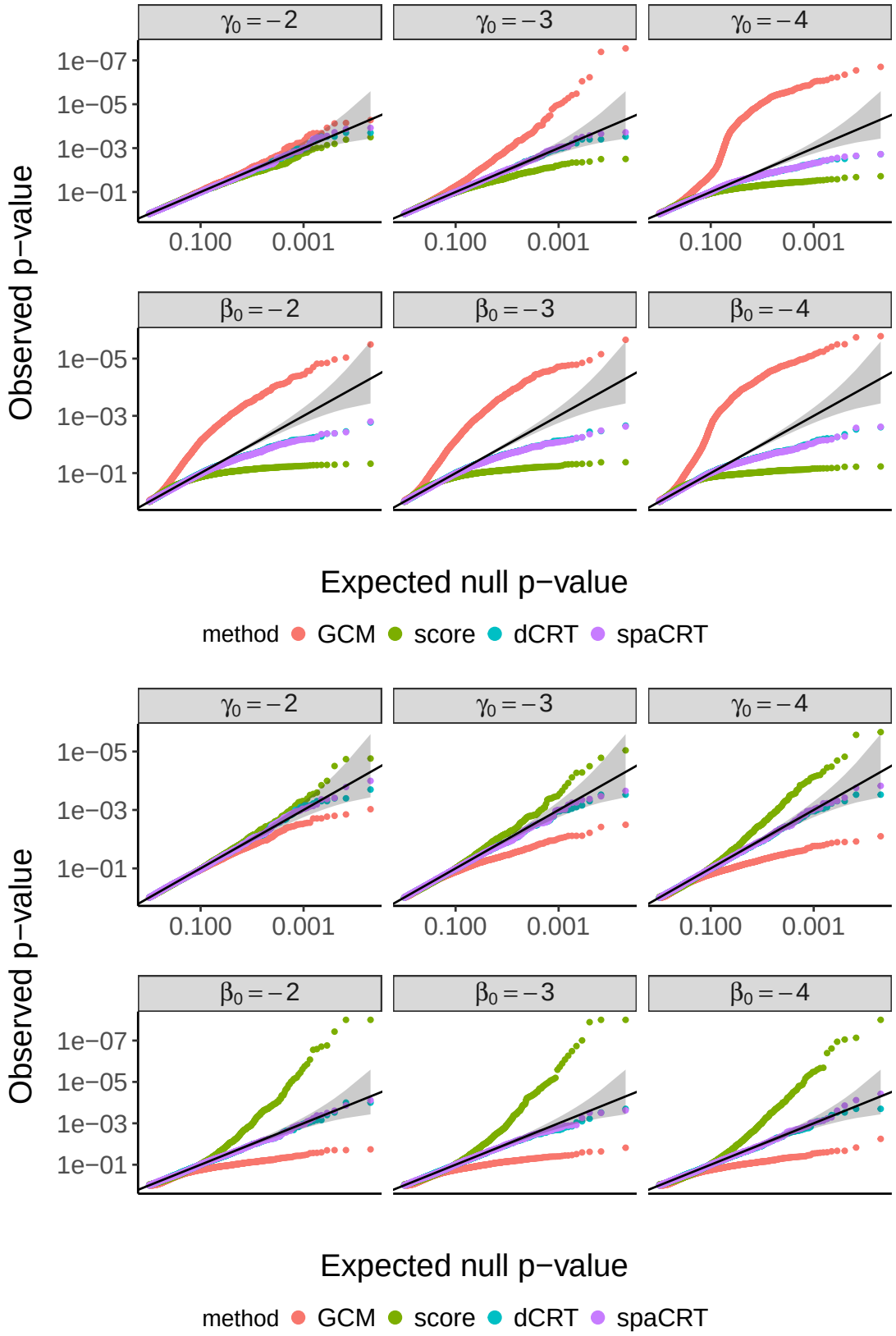

Figure 9: QQ plots of the p-values when $r = 0.05$. Top: left-sided test. Bottom: right-sided test.

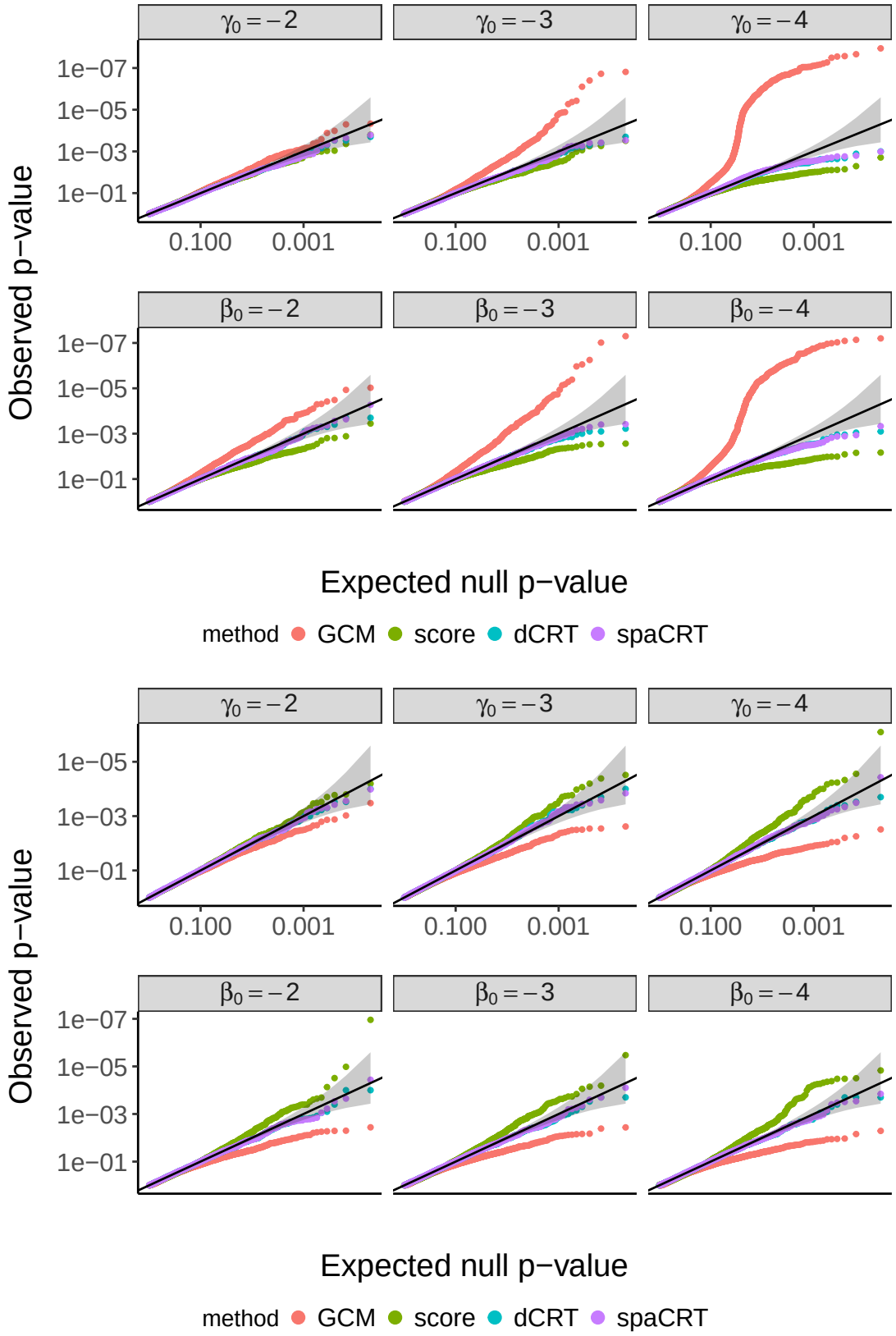

Figure 10: QQ plots of the p-values when $r = 1$. Top: left-sided test. Bottom: right-sided test.

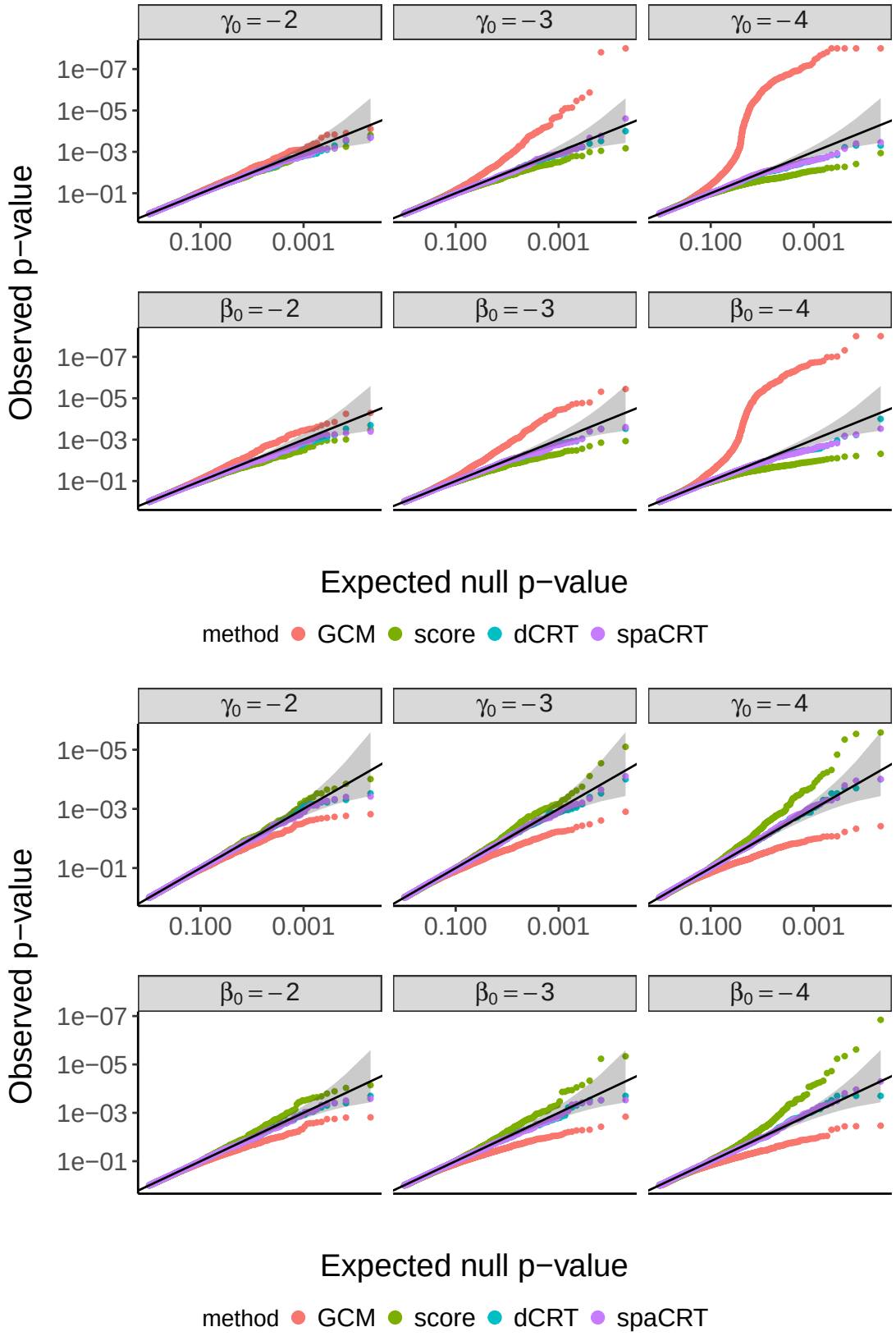

Figure 11: QQ plots of the p-values when $r = 10$. Top: left-sided test. Bottom: right-sided test.

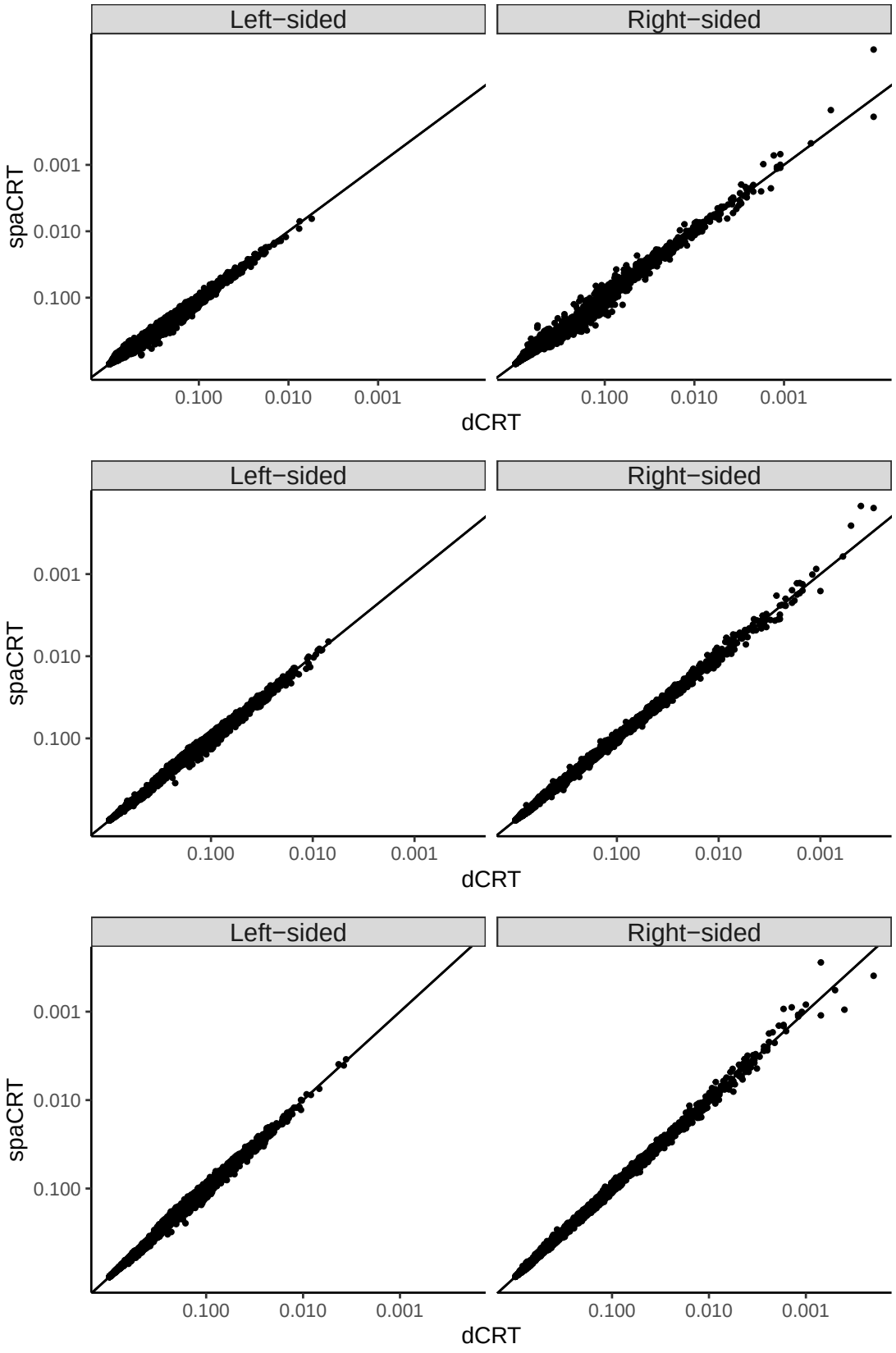

Figure 12: p-value approximation accuracy of spaCRT and dCRT when $r = 0.05, 1, 10$. Top: $r = 0.05$. Middle: $r = 1$. Bottom: $r = 10$.

Q Additional simulation details in Section 5.2

Q.1 Source of sparsity in GWAS with rare phenotypes

GWAS aim 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.

Q.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 consider $\boldsymbol{U}_j \in \{1, 2, \ldots, 10\}$ and $\boldsymbol{X} \in \{0, 1\}$. Therefore $K = 10$ and $M = 2$. 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:

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

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 choice will induce low sparsity in X and thus mimic the common genetic variation scenario. The simulation setup can be summarized in Table 4.

We will consider regularization parameters λ used in `glmnet` to be $\lambda =$ `lambda.1se` or $\lambda =$ `lambda.min`.

Table 4: 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 for Section 5.2.

- The **spaCRT** (Algorithm 2), where the parameters of the HMM are fitted based on expectation-maximization (EM) algorithm implemented by `fastPhase` software (Scheet, Stephens, and Scheet, 2006). `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), using the tower trick. We refer the details of methods implementation to Appendix Q.3.

- The **Knockoffs** (Barber and Candés, 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.

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

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 $j \in [d]$ 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, Stephens, and Scheet, 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.

Q.4 Additional simulation results in Section 5.2

We present additional simulation results in this section. The results can be found in Figure 13, 14 and 15. We want to point out that the choice of regularization parameter λ seems to affect both the FDR and power of the methods. In particular, comparing Figure 13 and Figure 14, 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.

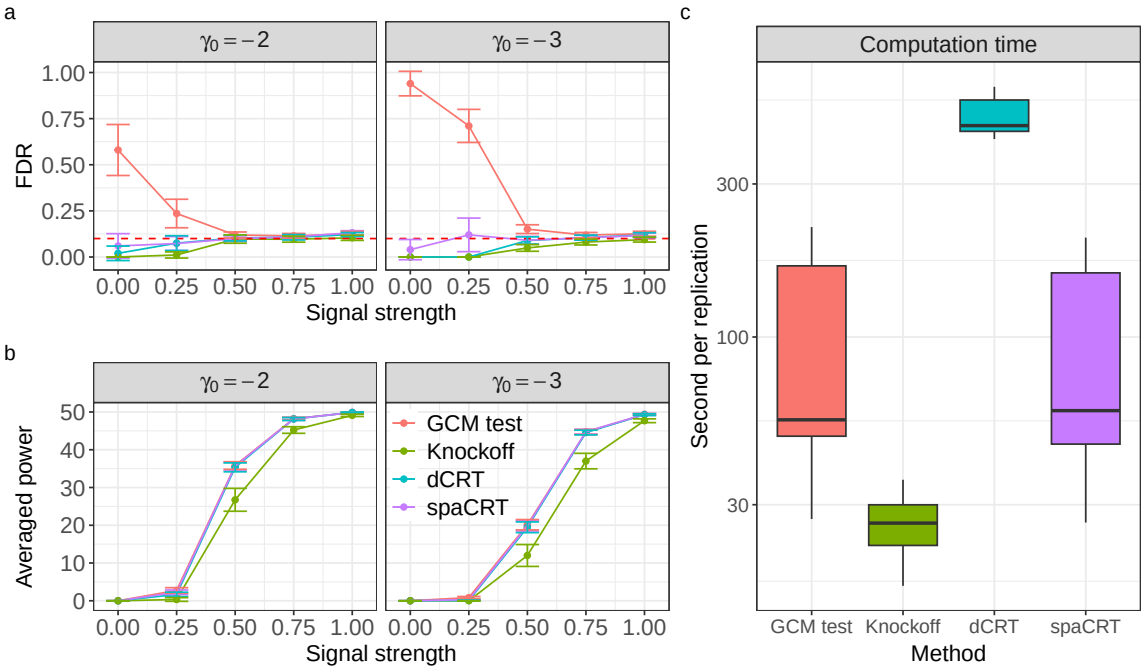

Figure 13: Summary of numerical simulation results for variable selection with $(\alpha, \beta) = (1, 1)$ and $\lambda =$ `lambda.1se`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times by different methods.

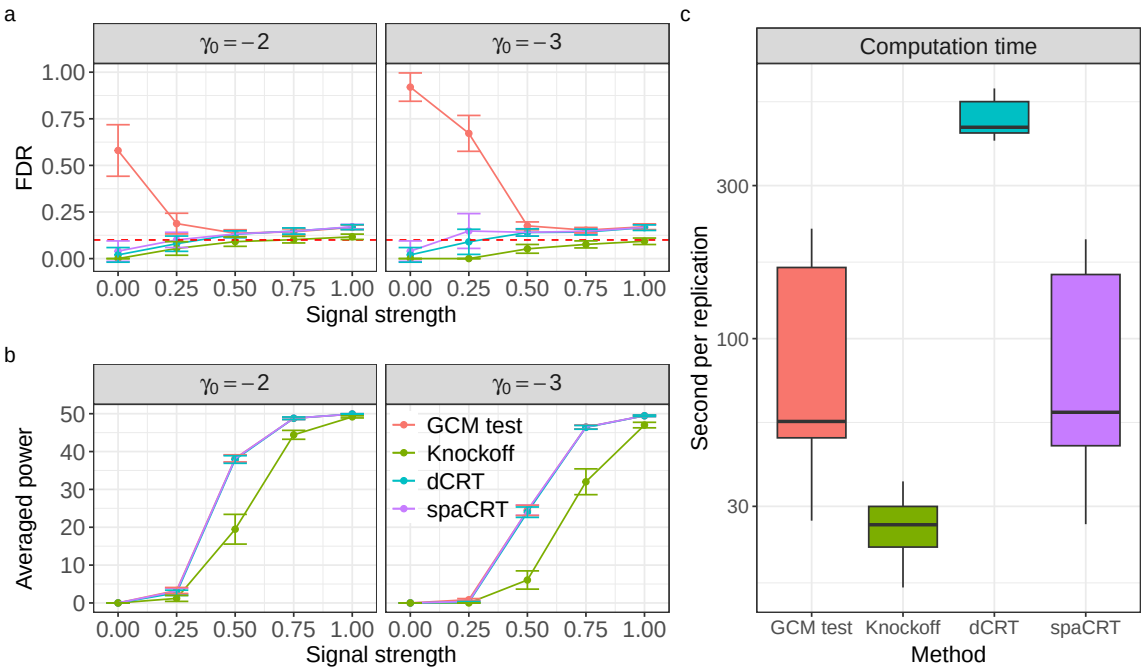

Figure 14: Summary of numerical simulation results for variable selection with $(\alpha, \beta) = (1, 1)$ and $\lambda =$ `lambda.min`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times by different methods.

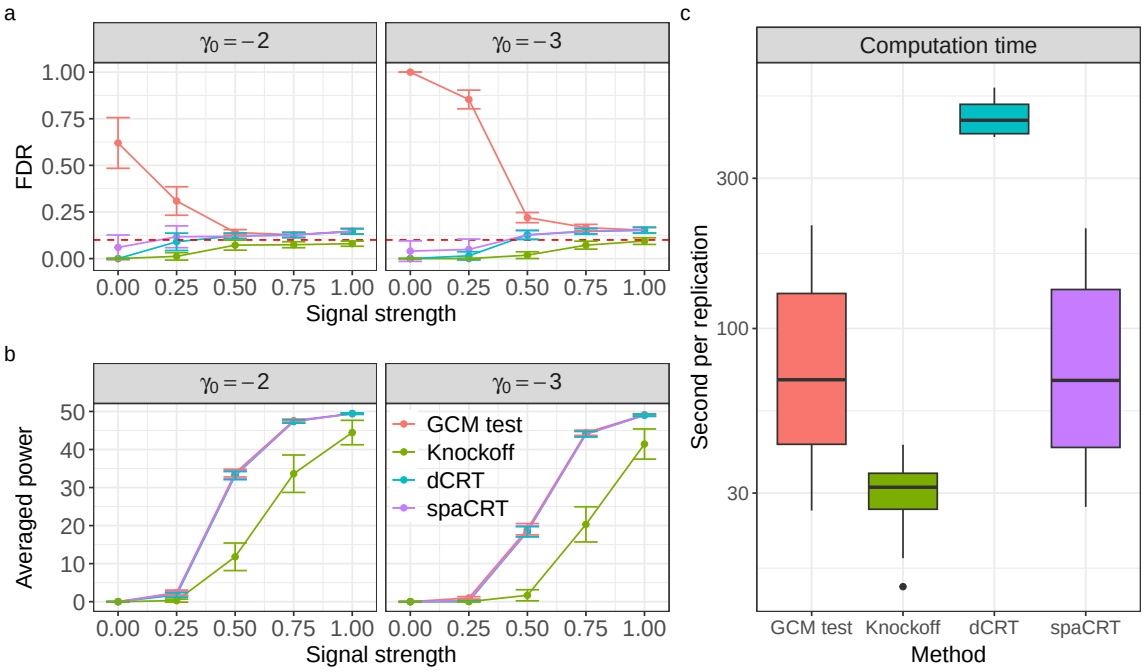

Figure 15: Summary of numerical simulation results for variable selection with $(\alpha, \beta) = (1, 3)$ and $\lambda =$ `lambda.min`. (a) FDR for $\gamma_0 = -3$ (high sparsity) and $\gamma_0 = -2$ (low sparsity). (b) Power for the same set of γ_0. (c) Computation times by different methods.

R Additional figures and tables for real data analysis

We present relevant information about Gasperinit dataset in Section R.1. In Section R.3, 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 R.4, we present additional figures for real data analysis including QQ-plots faceting across different effective sample size (Figure 17) and QQ-plots faceting across different dispersion parameters (Figure 18). We report the failure of spaCRT occurring in at most 0.007% of all hypotheses tested.

R.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.

R.2 Data sparsity

To demonstrate that the sparsity in the real data, Figure 16 summarizes the sparsity of gRNA presence and gene expression in our real single-cell CRISPR dataset. Panel a shows that many genes exhibit expression rates near 0.01, while Panel b shows that gRNA perturbation presence are generally even lower. We also show the effective sample size of the 153,000 negative control pairs in Table 5. We comment that the sparsity in real data roughly matches what we chose in the simulation (see Section P.2). The effective sample size is defined as the number of cells in which both the gRNA and gene expression are non-zero.

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

Table 5: Effective sample sizes in the 153,000 negative control pairs.

R.3 Additional table for the real data analysis

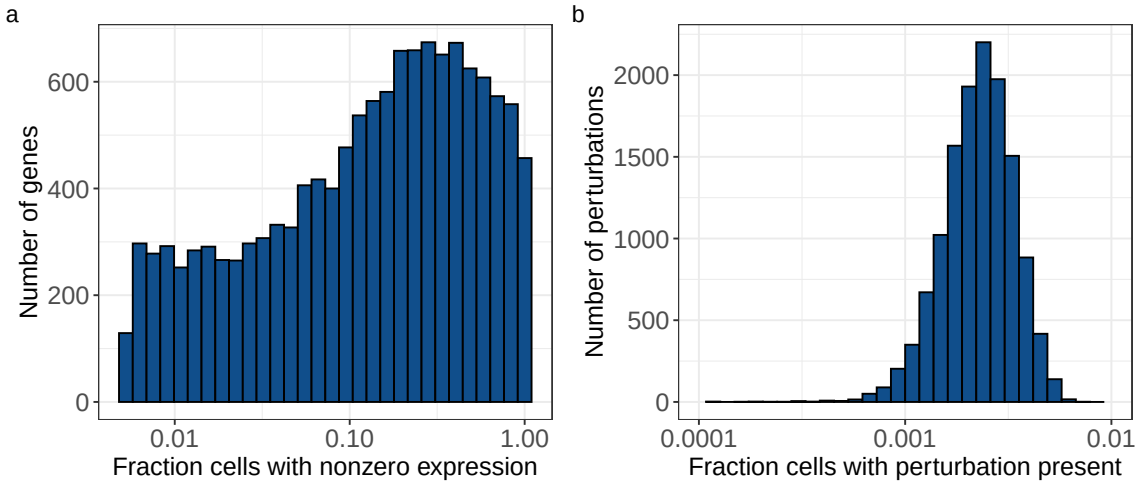

Figure 16: Histograms of the sparsity in Gasperini et al. (2019). (a) The histogram of the sparsity of gene expression. (b) The histogram of the sparsity of gRNA presence.

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

R.4 Additional figures for the real data analysis

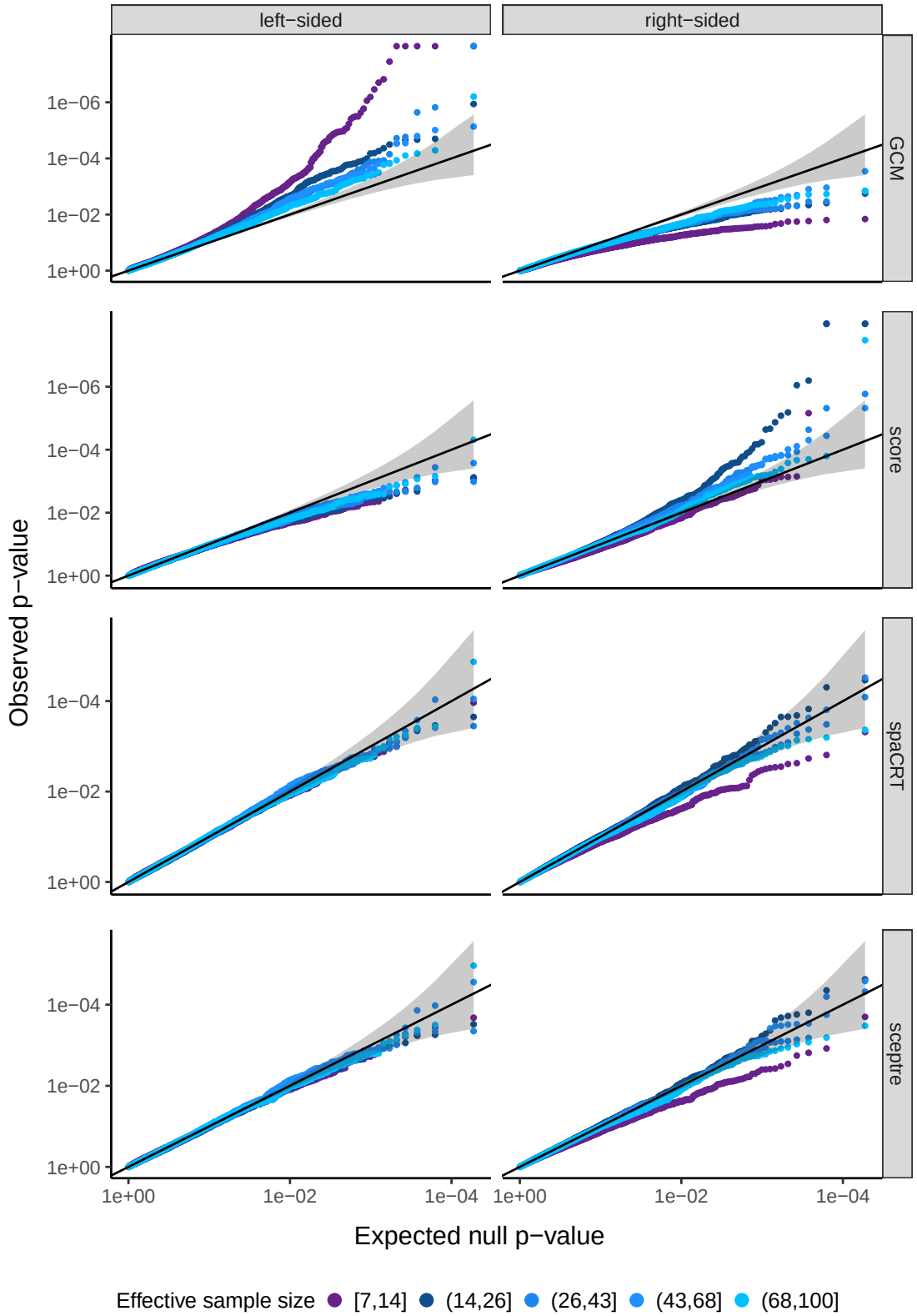

Figure 17: QQ-plots for the p-values of right-sided test from different methods under low effective sample size.

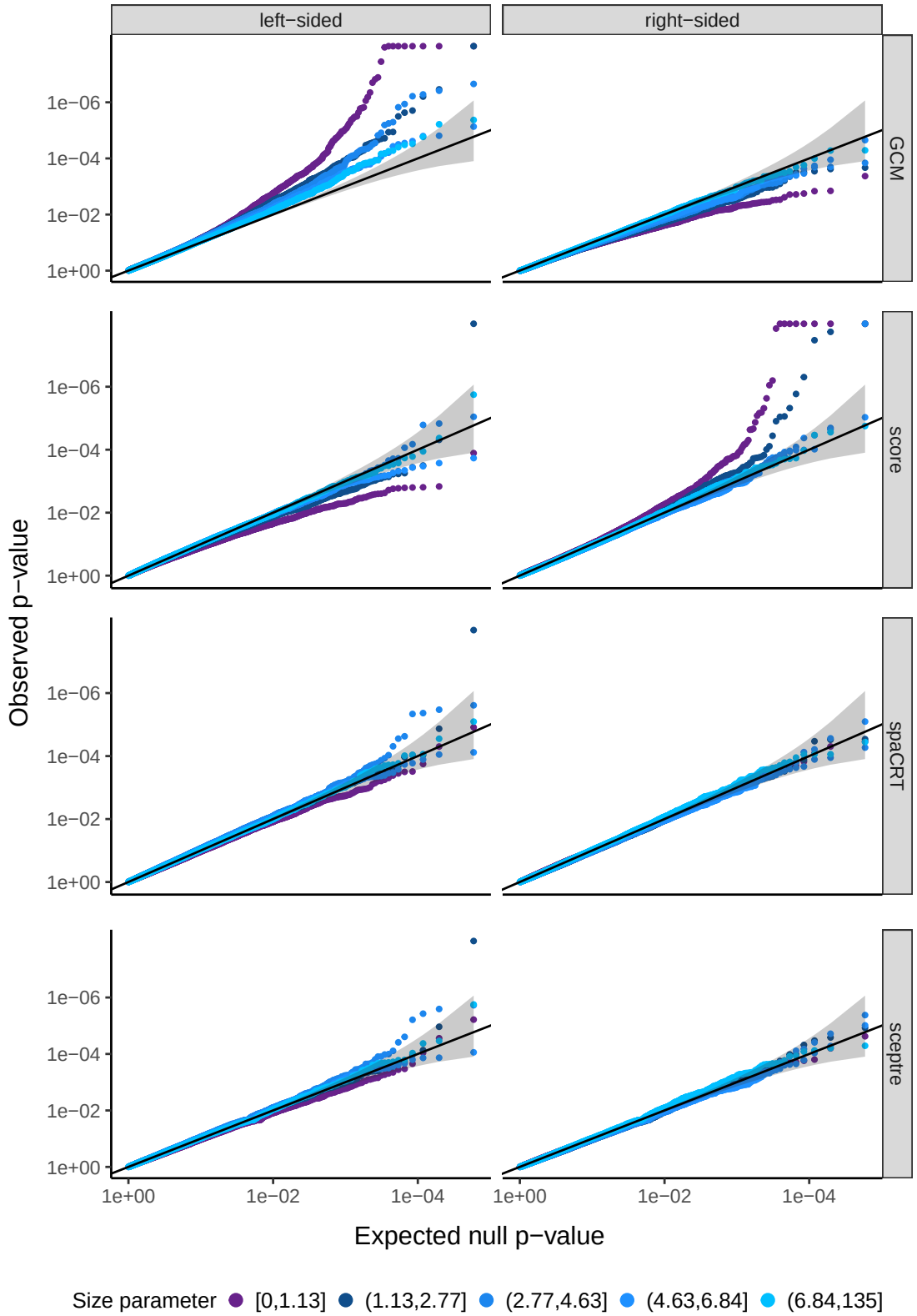

Figure 18: QQ-plots for the p-values of left-sided test from different methods stratified by dispersion parameter.