A Power Analysis of the Conditional Randomization Test and Knockoffs

Wenshuo Wang and Lucas Janson

Department of Statistics, Harvard University

Abstract

In many scientific problems, researchers try to relate a response variable Y to a set of potential explanatory variables $X = (X_1, \ldots, X_p)$, and start by trying to identify variables that contribute to this relationship. In statistical terms, this goal can be posed as trying to identify X_j's upon which Y is conditionally dependent. Sometimes it is of value to simultaneously test for each j, which is more commonly known as variable selection. The conditional randomization test (CRT) and model-X knockoffs are two recently proposed methods that respectively perform conditional independence testing and variable selection by, for each X_j, computing any test statistic on the data and assessing that test statistic's significance by comparing it to test statistics computed on synthetic variables generated using knowledge of X's distribution. Our main contribution is to analyze their power in a high-dimensional linear model where the ratio of the dimension p and the sample size n converge to a positive constant. We give explicit expressions of the asymptotic power of the CRT, variable selection with CRT p-values, and model-X knockoffs, each with a test statistic based on either the marginal covariance, the least squares coefficient, or the lasso. One useful application of our analysis is the direct theoretical comparison of the asymptotic powers of variable selection with CRT p-values and model-X knockoffs; in the instances with independent covariates that we consider, the CRT provably dominates knockoffs. We also analyze the power gain from using unlabeled data in the CRT when limited knowledge of X's distribution is available, and the power of the CRT when samples are collected retrospectively.

Keywords. Conditional randomization testing, knockoffs, Benjamini–Hochberg, model-X, retrospective sampling, approximate message passing.

1 Introduction

1.1 The conditional randomization test and model-X knockoffs

Analyzing the statistical relationship between random variables lies at the heart of many practical problems. For example, in clinical trials, doctors aim to determine whether a certain treatment has any effect on the patient's health. In genome-wide association studies (GWAS), researchers seek to understand which genes directly contribute to a trait of interest. Many such modern statistical problems are set up in high dimensions, partly because scientific advances have allowed us to easily collect a large number of covariates.

Candès et al. (2018) proposed two methods to identify important variables: the conditional randomization test (CRT) for testing conditional independence, and model-X knockoffs, or simply knockoffs, for variable selection while controlling the false discovery rate (FDR). Coined in Candès

et al. (2018), "model-X" refers to a framework where inference is conducted by making as many assumptions on the distribution of X (covariates) as possible and as few assumptions on the conditional distribution of Y (outcome) given X as possible. While the CRT and knockoffs have gained the interest of many researchers, there has been limited theoretical work on their power.

1.2 Our contribution

This article analyzes the CRT for single hypothesis testing, its generalization for multiple hypothesis testing, and knockoffs for variable selection (Candès et al. 2018). We mainly study the asymptotic performance of the CRT and knockoffs with different choices of popular test statistics. Power analysis is beneficial for various reasons. First, it is useful for determining how many samples one would like to collect beforehand in order to achieve a certain target power in a given experiment. Second, it allows direct comparison between methods in infinitely many data-generating distributions without the need for any simulations. For CRT and knockoffs, power analysis is particularly important for two reasons: (a) both methods act as wrappers around a chosen test statistic, and our theory can be used to choose among test statistics according to their power for a given data-generating distribution; (b) both methods provide particularly easy ways to leverage unlabeled data and also apply directly to retrospectively sampled data, though the impact of the unlabeled data or the retrospective sampling scheme on power has not been studied.

Our results are in the setting of high-dimensional linear regression, where the ratio of the numbers of observations and covariates converges to a positive constant and the covariates follow a multivariate Gaussian distribution $\mathcal{N}(0, \Sigma)$. Our main results are:

1. We give explicit expressions for the asymptotic power of the CRT when the test statistic is derived from the marginal covariance, ordinary least square (OLS) coefficient, or the lasso (Tibshirani 1996). We also prove bounds on and conjecture an exact expression for the asymptotic power of the CRT with marginal covariance test statistic when finite unlabeled samples are incorporated.

2. When $\Sigma = I$, i.e., the covariates are independent and identically distributed (i.i.d.) Gaussian random variables, we characterize the asymptotic power of the Benjamini–Hochberg (BH) procedure and the adaptive p-value thresholding (AdaPT) (Lei and Fithian 2018) procedure applied to the CRT p-values given by the aforementioned three statistics. In the same setting, we also show all of these procedures asymptotically control the FDR at the nominal level.

3. When $\Sigma = I$, we derive the asymptotic power of knockoffs using statistics derived from the marginal covariance, the OLS coefficient, or the lasso. We show that knockoffs is asymptotically equivalent to applying the AdaPT procedure to CDF-transformed knockoff statistics, and thus we can directly compare knockoffs' power with our earlier results on the CRT.

4. We extend the above three contributions with the marginal covariance test statistic to retrospectively sampled data, showing that the resulting effective signal strength is an explicit function of the marginal second moment of the retrospectively sampled Y.

We demonstrate that our asymptotic power expressions are quite accurate in finite samples, and the CRT and knockoffs can achieve power close to oracle methods.

1.3 Related work

Since Candès et al. (2018) introduced the CRT and model-X knockoffs, subsequent works have studied their robustness (Barber et al. 2018; Berrett et al. 2019; Huang and Janson 2020; Barber

and Candès 2019), computation (Tansey et al. 2018; Bates et al. 2020a; Liu et al. 2020), and application to, e.g., neural networks (Lu et al. 2018), time series (Fan et al. 2018), graphical models (Li and Maathuis 2019), and biology (Sesia et al. 2018; Katsevich and Sabatti 2019; Sesia et al. 2020b; Bates et al. 2020b; Sesia et al. 2020a; Chia et al. 2020; Katsevich and Roeder 2020).

Regarding the power of these methods, Weinstein et al. (2017) analyzed the power of a knockoffs-inspired procedure that is only valid when all the covariates are i.i.d.; our work studies (in addition to the CRT) the original model-X knockoffs procedure, which is valid for any covariate distribution although we study it in a setting with i.i.d. covariates. Liu and Rigollet (2019) provided a condition under which knockoffs' power goes to 1 and FDR goes to 0; our work exactly characterizes the asymptotic power when it is non-trivial (strictly between 0 and 1). Katsevich and Ramdas (2020) studied the CRT under low-dimensional asymptotics, while our work focuses on the high-dimensional regime, although we include a short note on the power of the CRT in low dimensions in Appendix B. During the preparation of our manuscript, Weinstein et al. (2020) independently quantified the asymptotic power of knockoffs with the lasso coefficient difference statistic, a result which is quite similar to our Theorem 7, though that paper does not study the CRT or other statistics for knockoffs as we do.

There have been a number of other works on the asymptotic power of other methods that test for covariate importance (Zhu and Bradic 2018; Chernozhukov et al. 2018; Javanmard et al. 2018). These methods are fundamentally different from the CRT and knockoffs, but we will compare their results with our own in Section 2.2.5.

1.4 Notation

Bold letters are used for matrices or vectors containing i.i.d. observations. For a set S, $|S|$ denotes the number of elements in S. The cumulative distribution function (CDF) of the Gaussian distribution $\mathcal{N}(0,1)$ is denoted by Φ—for $\alpha \in (0,1)$, z_α denotes the α-quantile of $\mathcal{N}(0,1)$, i.e., $\Phi(z_\alpha) = \alpha$. For random variables or vectors W_1 and W_2, $\mathcal{L}(W_1)$ means the distribution of W_1 and $\mathcal{L}(W_1 \,|\, W_2)$ means the conditional distribution of W_1 given W_2.

1.5 Outline of the article

In Section 2, we analyze the CRT's power for single hypothesis testing (conditional independence testing), including the case where we leverage unlabeled samples. In Section 3, we analyze the power of the CRT and knockoffs for multiple testing (variable selection). In Section 4, we study the effect of retrospective sampling. Section 5 supports our theoretical results with simulations. Finally, we conclude with a discussion of some questions raised by our work in Section 6.

2 Power analysis of conditional independence testing

In this section, we study the power of the CRT for testing a single hypothesis of conditional independence.

2.1 The conditional randomization test

We begin with a review of the CRT introduced in Candès et al. (2018). Consider the generic problem of testing $H_0^{(j)} : X_j \perp\!\!\!\perp Y \mid X_{-j}$ in a regression setting where we have n i.i.d. observations. To lighten notation, we label X_j as simply X and X_{-j} as Z, and the data matrix is therefore denoted by $[\mathbf{X}, \mathbf{Z}, \mathbf{Y}]$, where $\mathbf{X} \in \mathbb{R}^n$ is the covariate vector of interest, $\mathbf{Z} \in \mathbb{R}^{n \times (p-1)}$ is the matrix

of confounding variables, and $\mathbf{Y} \in \mathbb{R}^n$ is the response vector. Suppose we have a test statistic function T of $(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ that intuitively measures the importance of variable X (e.g., T could be the absolute value of the coefficient for $\mathbf{X}$ fitted by the lasso). To construct a test, we need to find a cutoff for the test statistic $T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ such that we can guarantee $T(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$ only falls above that cutoff with probability at most the nominal level α under $H_0^{(j)}$. This requires some knowledge of its distribution under the null. Candès et al. (2018) suggested the following: if we know $\mathcal{L}(X \mid Z)$, then let $\tilde{\mathbf{X}} \mid \mathbf{Z}, \mathbf{Y} \sim \mathcal{L}(\mathbf{X} \mid \mathbf{Z})$ and we will have

$$T(\mathbf{X}, \mathbf{Z}, \mathbf{Y}) \stackrel{d}{=} T(\tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{Y}) \mid \mathbf{Z}, \mathbf{Y} \text{ under the null.}$$

Thus, we can obtain a cutoff using the known distribution $\mathcal{L}(T(\tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{Y}) \mid \mathbf{Z}, \mathbf{Y})$. Although such a cutoff can only be computed analytically in special cases (Liu et al. 2020), an empirical one can be obtained by repeatedly resampling $\tilde{\mathbf{X}}$ and recomputing $T(\tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{Y})$. The CRT with an empirical cutoff contains a finite-sample correction to make the test exact, but it converges to the test using the exact quantile if the number of resamples M_n goes to infinity. The cases we consider in this article all have an analytical cutoff available and this is the CRT we study, but the same results would still hold as long as $M_n \to \infty$. It is worthwhile to emphasize that *any* test statistic function T can be used in the CRT.

We have assumed that we know $\mathcal{L}(X \mid Z)$ exactly, and will make this assumption in Section 2.2; in Section 2.3, we will study the power when this assumption is relaxed by conditioning and leveraging unlabeled data (Huang and Janson 2020), and in Section 4, we will discuss how the power changes with retrospective sampling (Barber and Candès 2019).

2.2 CRT in high-dimensional linear regression

We begin by analyzing the power of the CRT using several different statistics in the high-dimensional linear regression setting formally defined as follows in Setting 1.

Setting 1 (High-dimensional linear model)**.** *Consider the linear regression model*

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $\mathbf{X} \in \mathbb{R}^n$*,* $\mathbf{Z} \in \mathbb{R}^{n \times (p-1)}$*, and for each row, they satisfy*

$$\mathbf{Z}_{i\cdot} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_Z), \qquad \mathbf{X}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{Z}_{i\cdot}^\top \xi, 1), \qquad (\mathbf{X}, \mathbf{Z}) \perp\!\!\!\perp \varepsilon.$$

This setting assumes the above model under the following high-dimensional asymptotics:

$$\lim_{n\to\infty} p/n \to \kappa \in (0, \infty), \qquad \lim_{n\to\infty} \theta^\top \Sigma_Z \theta \to v_Z^2, \qquad \limsup_{n\to\infty} \xi^\top \Sigma_Z \xi < \infty,$$

and σ^2 *and* $\sqrt{n}\beta = h > 0$ *stay constant.*

We emphasize again that the assumptions in Setting 1 are for the study of power and are not needed for the validity of the CRT. Here, $\theta^\top \Sigma_Z \theta$ can be interpreted as the part of Y's variance contributed by Z, as $\theta^\top \Sigma_Z \theta = \mathrm{Var}(\mathbb{E}[Y \mid Z])$; similarly, $\xi^\top \Sigma_Z \xi$ can be interpreted as the part of X's variance contributed by Z. For instance, the assumptions on $\theta^\top \Sigma_Z \theta$ and $\xi^\top \Sigma_Z \xi$ hold if $\Sigma_Z = I$ and the components of θ and ξ are $n^{-1/2}$-normalized i.i.d. draws from a distribution with finite second moment. More concretely, if $\sqrt{n}\theta_j \overset{\text{i.i.d.}}{\sim} B_0$ and $\Sigma_Z = I$, then $\theta^\top \Sigma_Z \theta \to \kappa \mathbb{E}[B_0^2]$ almost surely, which corresponds to the setting we will consider in Section 3.

The CRT tests $H_0 : X \perp\!\!\!\perp Y \mid Z$, which, under Setting 1, is equivalent to $H_0 : \beta = 0$, and we are interested in the power under local alternatives $H_1 : \beta = h/\sqrt{n}$ for a fixed $h > 0$, which is the regime where the power has a non-trivial limit strictly between 0 and 1. In this section, the asymptotic power means the limit of the unconditional power of the test under $\beta = h/\sqrt{n}$. We consider three different test statistics for the CRT and for each one we will show there exists a scalar μ (which we will give an expression for) such that the asymptotic power is equal to that of a z-test with standardized effect size μ as in Definition 1.

Definition 1. *The CRT with test statistic T under a given asymptotic regime is said to have asymptotic power equal to that of a z-test with standardized effect size μ, if the level-α one-sided CRT (reject for large values of T) has asymptotic power $\Phi(\mu - z_{1-\alpha})$ and the level-α two-sided CRT (reject for large values of $|T|$) has asymptotic power $\Phi(\mu - z_{1-\alpha/2}) + \Phi(-\mu - z_{1-\alpha/2})$.*

2.2.1 Marginal covariance

Consider testing using the statistic $T_{\mathrm{MC}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = n^{-1}\mathbf{Y}^\top \mathbf{X}$, which is an unbiased and consistent estimator of $\mathrm{Cov}(X, Y)$. Although it may seem naïve to consider a marginal test statistic that does not involve Z, it is actually a popular choice in many high-dimensional applications such as genome-wide association studies (Wu et al. 2010) and microbiome studies (McMurdie and Holmes 2014). T_{MC} is simple and intuitive and we will see it performs well when the confounding vector Z does not contribute too much variance to Y.

Theorem 1. *In Setting 1, the CRT with T_{MC} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\sqrt{\sigma^2 + v_Z^2}}.$$

We first note that the power increases as h increases, which is intuitive because h is the coefficient (dropping the normalization term $\sqrt{n}$) and the effective signal strength. Second, the dimensionality (or equivalently, κ) does not appear explicitly, which can be understood by noticing that Z only plays a role through $Z^\top\theta$, which we can consider as part of the error, thus adding extra variance $\theta^\top \Sigma_Z \theta \to v_Z^2$. Then the asymptotic effective error variance is $\sigma^2 + v_Z^2$, and when this number is large, the power is low.

2.2.2 Ordinary least squares

There are many reasons why one might want to use the ordinary least squares (OLS) estimate $T_{\mathrm{OLS}} = \hat{\beta}^{\mathrm{OLS}}$ as the test statistic; for instance, it is the maximum likelihood estimator and the best linear unbiased estimator. In this section, we will assume $\kappa < 1$ in Setting 1 so that $\hat{\beta}^{\mathrm{OLS}}$ is well-defined.

Theorem 2. *In Setting 1 with $\kappa < 1$, the CRT with T_{OLS} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\sigma}\sqrt{1-\kappa}.$$

We can see that the power decreases as σ and κ increase. Compared to using T_{MC}, the CRT with T_{OLS} has higher power if $\kappa < v_Z^2/(\sigma^2 + v_Z^2)$, and vice versa. In fact, as κ approach 1, OLS becomes

ill-defined and the test becomes powerless. As another comparison, consider the canonical OLS test that takes $\hat{\beta}^{\text{OLS}}$ as the test statistic and conducts a test based on the conditional distribution $\hat{\beta}^{\text{OLS}} \mid \mathbf{X}, \mathbf{Z} \sim \mathcal{N}\left(\beta, \sigma^2 \left([\mathbf{X}, \mathbf{Z}]^\top [\mathbf{X}, \mathbf{Z}]\right)^{-1}_{11}\right)$. This canonical test turns out to have the same asymptotic power as the one for the CRT given in Theorem 2 (see Appendix F for derivation).

2.2.3 The distilled lasso statistic

For high-dimensional regression, one might naturally look to the lasso to construct a test statistic. In this section, we consider the distilled lasso statistic, a test statistic proposed in Liu et al. (2020) derived from the lasso, which leverages the lasso for fitting a high-dimensional coefficient vector and has very similar power to, but is more computationally efficient than, using $\hat{\beta}^{\text{lasso}}$ as the test statistic. In our notation, the statistic can be defined as follows. We first regress $\mathbf{Y}$ on $\mathbf{Z}$ using the lasso with penalty parameter λ to obtain $\hat{\theta}_\lambda$, i.e.,

$$\hat{\theta}_\lambda = \underset{\theta}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{Z}\theta\|_2^2 + \sqrt{n}\lambda\|\theta\|_1.$$

The distilled lasso statistic is then defined as $T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = n^{-1}(\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda)^\top(\mathbf{X} - \mathbf{Z}\xi)$. Intuitively, it measures the covariance of X and Y after their dependence on Z is removed, where Y's dependence on Z is estimated by the lasso.

To analyze this test statistic, we will leverage the theory of approximate message passing (AMP), which has been used to characterize the asymptotic distribution of the lasso coefficient vector (Bayati and Montanari 2011). This asymptotic distribution depends on two important parameters α_λ and τ_λ which are uniquely defined as the solution to a system of explicit fixed-point equations depending on λ, σ^2, κ, and the asymptotic histogram of the true coefficients $\sqrt{n}\theta_j$'s.[1] In Appendix E, we provide the fixed-point equations (12). Intuitively speaking, $\sqrt{n}\hat{\theta}_j$ is roughly distributed as $\eta(\sqrt{n}\theta_j + \tau_\lambda Z; \alpha_\lambda\tau_\lambda)$, where

$$\eta(x; y) = \begin{cases} x - y, & x > y, \\ 0, & |x| \le y, \\ x + y, & x < -y, \end{cases} \tag{1}$$

$Z \sim \mathcal{N}(0, 1)$, so that τ_λ plays the role of the level of the asymptotic estimation error and α_λ acts as a soft-thresholding parameter. AMP theory, and hence our use of it, relies on additional assumptions as stated in Theorem 3.

Theorem 3. *Under Setting 1 with $\Sigma_Z = I$ and $\xi = 0$, if the empirical distribution of $(\sqrt{n}\theta_j)_{j=1}^{p-1}$ converges to a distribution represented by a random variable B_0 and $\|\sqrt{n}\theta\|_2^2/p \to \mathbb{E}[B_0^2]$, then the CRT with the distilled lasso statistic with lasso parameter λ has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\tau_\lambda}.$$

We prove Theorem 3 using results from Bayati and Montanari (2011). The key step is to analyze the asymptotic correlation between the errors and the fitted residuals of the lasso regression, which has not been studied before. Theorem 3 is clean in that it only depends on the model parameters through a scalar τ_λ, making it helpful for guiding the choice of λ.

[1] Note that α_λ is unrelated to the level α of the statistical test.

2.2.4 Comparison of CRT statistics

In Figure 1, we plot the relationship of the asymptotic power of the CRT with the distilled lasso statistic and λ in different settings and compare with that of the CRT using marginal covariance and OLS. We can see that the dependence of the power on λ is mild, and the distilled lasso statistic with a good λ is always better than marginal covariance and OLS in the considered parameter settings. This is not a coincidence: the best distilled lasso statistic has at least the same power as the marginal covariance and the OLS coefficient. To see this, note that if $\lambda = \infty$, $\hat{\theta}_\lambda = 0$ and $T_{\text{distilled}} = T_{\text{MC}}$; if $\lambda = 0$, $\hat{\theta}_\lambda = \hat{\theta}^{\text{OLS}} = (\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{Y}$, and $T_{\text{distilled}}$ is equal to the numerator of the expression of T_{OLS} in Equation (9) in the proof of Theorem 2, the power of which can be even more easily proved to be the same as T_{OLS} following the same proof.

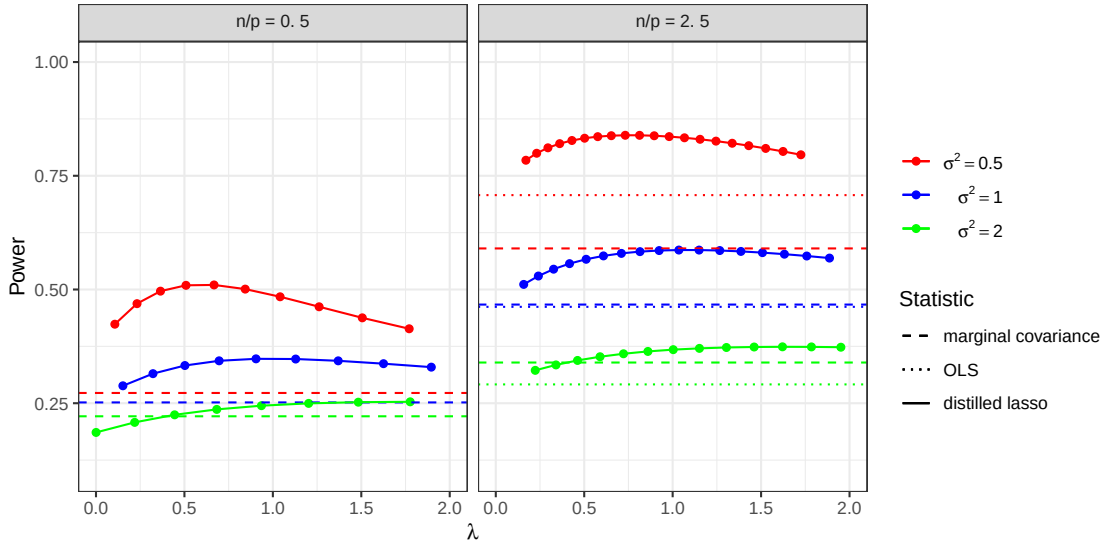

Figure 1: Dependence of the asymptotic power of the CRT with the distilled lasso statistic on λ, compared with that with the marginal covariance statistic and the OLS statistic; plots are in Setting 1 with $B_0 \sim 0.9\delta_0 + 0.1\delta_4$, $\Sigma_Z = I$ and $\xi = 0$, so $v_Z^2 = 1.6\kappa$. Signal size is $\beta = 2/\sqrt{n}$. OLS is not applicable for $n = 0.5p$.

2.2.5 Comparison with other methods

It is possible to achieve the power $\Phi\left(h/\sigma - z_\alpha\right)$ if θ and ξ can be estimated at a sufficiently fast rate, which dominates all three of our statistics (for the lasso, note that $\tau_\lambda \geq \sigma$). Javanmard et al. (2018) achieved this rate (their Theorem 3.8) by assuming, among other conditions, $\mathcal{L}(X, Z)$ is known and θ has sparsity level $o(n/(\log p)^2)$ (which is not satisfied in our setting). Chernozhukov et al. (2018) also obtained this rate (their Theorem 4.1) by assuming $\hat{\xi}$ and $\hat{\theta}$ are consistent and $\|\hat{\xi} - \xi\|\|\hat{\theta} - \theta\| = o(n^{-1/2})$, while it is well-known that consistency in high-dimensional settings is usually impossible without strong assumptions such as sparsity (which we do not make).

Zhu and Bradic (2018) assumed sub-Gaussianity and moment conditions to derive the same asymptotic power (their Theorem 7) as in our Theorem 1 for a different method they proposed and under different assumptions, except that there a two-sided test was considered. There are differences that are worth noting: (a) Zhu and Bradic (2018) does not assume ξ is known, but requires ξ to have sparsity $o(\sqrt{n}/(\log(\max(p, n)))^{5/2})$ in order to estimate it fast enough and gives an asymptotically valid test, and here we assume we know ξ so the test has exact size α for any finite (n, p); (b)

we make stronger assumptions on $\mathcal{L}(Y \mid X, Z)$, which is mainly to facilitate analysis for other more complex statistics; our Theorem 1 could easily be extended to only assume moment conditions on ε.

2.3 Leveraging unlabeled data in the CRT

In Section 2.2, we assumed we knew $\mathcal{L}(X \mid Z)$ exactly, which can be understood as a case in which we have sufficient unlabeled data and/or domain knowledge that we effectively know this distribution exactly. In some practical cases, however, we do not know $\mathcal{L}(X \mid Z)$ exactly and would like to leverage finite unlabeled samples to learn more about it. To this end, we explore in this section a useful idea introduced in Huang and Janson (2020), which only assumes a model on $\mathcal{L}(X \mid Z)$ and conditions on a sufficient statistic that uses both labeled and unlabeled data.

As a concrete example, we again consider Setting 1, but with the following changes: ξ is unknown; $\mathrm{Var}(X \mid Z) = 1$ but is unknown to the CRT. Effectively, we have assumed an unknown Gaussian linear model for $\mathcal{L}(X \mid Z)$. In this case, we would not be able to sample from $\mathcal{L}(\mathbf{X} \mid \mathbf{Z})$ to get $\tilde{\mathbf{X}}$ as normally required by the CRT. Exploiting Gaussianity, we can proceed by conditioning on a sufficient statistic as follows. Let $T(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$ be the test statistic and $S(\mathbf{X}, \mathbf{Z})$ be a sufficient statistic (e.g., $S(\mathbf{X}, \mathbf{Z}) = \mathbf{Z}^\top \mathbf{X}$ is sufficient in Setting 1) for the unknown parameter in $\mathcal{L}(\mathbf{X} \mid \mathbf{Z})$. By the sufficiency of S, $\mathcal{L}(\mathbf{X} \mid \mathbf{Z}, S(\mathbf{X}, \mathbf{Z}))$ does not depend on the unknown ξ or $\mathrm{Var}(X \mid Z)$, so we therefore know $\mathcal{L}(T(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Z}, S(\mathbf{X}, \mathbf{Z}), \mathbf{Y})$ under the null. Thus, we can obtain an analytical or empirical cutoff in the same way as the original unconditional CRT.

Although weakening the assumed knowledge of $\mathcal{L}(X \mid Z)$ by moving from an unconditional test to a conditional one may reduce the power, this reduction can be mitigated by incorporating unlabeled data into the sufficient statistic. Let the unlabeled samples be denoted by $(X_{n+i}, Z_{n+i})_{i=1}^m$, i.e., they are recorded without the response Y_{n+i}, where we assume m/p goes to a positive constant. Let $n_* = n+m$ and $\mathbf{X}_*$ and $\mathbf{Z}_*$ be the $n_* \times 1$ and $n_* \times (p-1)$ data matrices containing all X and Z samples, respectively, with the first n rows being labeled and corresponding to $\mathbf{Y}$. Similarly to the case without unlabeled data, let $T(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*)$ be the test statistic and $S(\mathbf{X}_*, \mathbf{Z}_*)$ be a sufficient statistic for the unknown parameter in $\mathcal{L}(\mathbf{X}_* \mid \mathbf{Z}_*)$, so that we know $\mathcal{L}(T(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*) \mid \mathbf{Z}_*, S(\mathbf{X}_*, \mathbf{Z}_*), \mathbf{Y})$ under the null. We emphasize that this idea also applies to non-Gaussian cases as long as a sufficient statistic exists. We can also see that $\mathcal{L}(\mathbf{X}_* \mid \mathbf{Z}_*, S(\mathbf{X}_*, \mathbf{Z}_*))$ does not change even if the labeled samples are collected retrospectively (i.e., based on the response variable Y; see, for example, Barber and Candès (2019)), as under the null, $\mathbf{X}_* \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}_*$. On the other hand, the power *will* change, though power analysis with both retrospective sampling and unlabeled data is beyond the scope of this article (while we do provide an analysis in Section 4 in the case of known $\mathcal{L}(X \mid Z)$, i.e., infinite unlabeled data); we focus on the case when the labeled samples are collected independent of the responses. It turns out that this procedure admits a quite substantial simplification for the Gaussian distribution. For instance, if T is chosen to be the marginal covariance $T_{\mathrm{MC}} = n^{-1}\mathbf{Y}^\top \mathbf{X}$, it simplifies to a statistic that could be seen as a generalization of the OLS statistic, which enables the analysis of its asymptotic power. We defer the details to Appendix G, where we also discuss why it might not be beneficial to consider the original OLS statistic in this setting. We present here upper and lower bounds of the asymptotic power together with a conjecture for its exact value.

Theorem 4. *In Setting 1 with ξ and $\mathrm{Var}(X \mid Z)$ unknown but fixed to be 1, if there are m additional data points $(X_i, Z_i)_{i=n+1}^{n+m}$, $n_* = n + m$, $n/n_* \to \kappa_*$ and $\kappa\kappa_* < 1$, then the conditional CRT with statistic T_{MC} has asymptotic power lower-bounded (the $\liminf$ is lower-bounded) by that of a z-test*

with standardized effect size

$$\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2\frac{1}{1-\kappa\kappa_*}}}$$

and upper-bounded (the lim sup *is upper-bounded) by that of a z-test with standardized effect size*

$$\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2\max\left(0,\frac{1-\frac{(1+\sqrt{1/\kappa})^2}{(1-\sqrt{\kappa\kappa_*})^2}\kappa\kappa_*}{1-\kappa\kappa_*}\right)}}.$$

Conjecture 1. *In Setting 1, if there are m additional data points $(X_i, Z_i)_{i=n+1}^{n+m}$, $n_* = n + m$, $n/n_* \to \kappa_*$ and $\kappa\kappa_* < 1$, then the conditional CRT with statistic T_{MC} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2(1-\kappa_*)}}.$$

See Figure 6 in Section 5 for a numerical validation. We discuss how we arrive at this conjecture in Appendix E. Note that the two bounds in Theorem 4 match the conjecture when $\kappa_* \to 0$.

Trivially when $\mathcal{L}(X \mid Z)$ is unknown, unlabeled data helps to run a test if $\kappa > 1$, since otherwise no non-trivial test can be run because the sufficient statistic uniquely determines $\mathbf{X}$'s exact value, making $\mathcal{L}(\mathbf{X} \mid \mathbf{Z}, S(\mathbf{X}, \mathbf{Z}))$ degenerate so that the only valid tests have power equal to their size under any alternative. When $\kappa < 1$, assuming Conjecture 1 holds, we see that unlabeled data can boost the power compared to using only labeled data ($\kappa_* = 1$) if $\kappa > v_Z^2/(\sigma^2 + v_Z^2)$, i.e., if p is close to n or if the nuisance variables Z contribute little variance to Y. This condition coincides with the condition under which the unconditional CRT with marginal covariance has higher power than with OLS. Another interesting takeaway is that if we keep $\kappa\kappa_*$ fixed and let $\kappa_* \to 0$, the asymptotic power is equal to that of a z-test with standardized effect size

$$\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2}}.$$

This can be interpreted as a setting where $p/n \to \infty$, but the number of unlabeled samples n_* scales with p as p/n_* goes to a non-zero constant $\kappa\kappa_*$.

3 Power analysis of variable selection

In this section, we consider variable selection and return to our original notation X_j and X_{-j} instead of X and Z (analogously for their bold counterparts), which were used in Section 2 in their place while j was fixed. More specifically, suppose we have a data matrix $[\mathbf{X}, \mathbf{Y}]$, where each row is an i.i.d. draw from a distribution $F_{X,Y}$, where X is a p-dimensional random vector and Y is a random variable. We can define variable selection as simultaneously testing the null hypotheses $H_0^{(1)}, \ldots, H_0^{(p)}$, where $H_0^{(j)}$ is $X_j \perp\!\!\!\perp Y \mid X_{-j}$. In this section, the power means the expectation of the ratio between the number of true discoveries and the number of non-null covariates.

To enable theoretical analysis, we study the linear regression setting with independent Gaussian covariates as given in Setting 2, where $H_0^{(j)}$ reduces to $\beta_j = 0$.

Setting 2 (High-dimensional linear model with independent covariates)**.** *Consider the linear regression model*

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I),$$

where $\mathbf{X} \in \mathbb{R}^{n\times p}$ *is a random matrix,*

$$X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \mathbf{X} \perp\!\!\!\perp \varepsilon.$$

This setting assumes the above model under the following high-dimensional asymptotics:

$$\lim_{n\to\infty} p/n = \kappa \in (0,\infty), \qquad \sqrt{n}\beta_j \overset{\text{i.i.d.}}{\sim} \gamma\delta_0 + (1-\gamma)\pi_1,$$

where γ *and* π_1 *are fixed,* π_1 *has bounded support and puts no mass at* 0*, and* $\beta \perp\!\!\!\perp (\mathbf{X}, \varepsilon)$*.*

In the future, we will use B_0 to represent a random variable following $\gamma\delta_0 + (1-\gamma)\pi_1$. Setting 2 is a slight modification of Setting 1 that makes all X_j's exchangeable. The Gaussian assumption makes theoretical derivation easier, and allows for the use of results on the lasso obtained by AMP theory. While the model is not believed to be appropriate if the covariates are too dependent on each other, in many applications the covariates are only slightly correlated. Hence, although a simple setting, Setting 2 is expected to be of value and can still guide statistic choice in many applications.

We analyze two types of procedures that control the FDR or asymptotic FDR in this setting.

1. BH and AdaPT applied to p-values obtained by the CRT. To the best of our knowledge, these are the first results on the validity or power of BH and AdaPT applied to CRT p-values.

2. The model-X knockoff filter (Candès et al. 2018).

3.1 Variable selection with the CRT

A natural way of generalizing the conditional independence tests of Section 2 to variable selection is to take the p-values from the CRT and plug them into a multiple testing procedure. Here, then, we consider the BH procedure (Benjamini and Hochberg 1995) and the AdaPT procedure (Lei and Fithian 2018) for controlling the FDR, defined as

$$\text{FDR} = \mathbb{E}\left[\text{FDP}\right], \qquad \text{FDP} = \frac{|\hat{S} \cap S_0|}{\max(|\hat{S}|, 1)},$$

where FDP stands for false discovery proportion, S_0 is the set of null variables and $\hat{S}$ is the set of selected variables. When we refer to AdaPT, we mean the intercept-only AdaPT procedure, (i.e., AdaPT without side information), which rejects all p-values below

$$\max\{t \in [0,1] : \frac{1 + \#\{j : p_j \geq 1-t\}}{\#\{j : p_j \leq t\}} \leq q\},$$

with q being the target FDR level. BH is the most used multiple testing procedure for controlling the FDR, and studying AdaPT allows us to directly compare variable selection using the CRT with knockoffs due to an asymptotic equivalence between knockoffs and a certain application of AdaPT, which we will explain in Section 3.2.2. It is known that BH and AdaPT control the FDR at the nominal level when all p-values are independent and the null p-values follow the standard uniform distribution on $[0,1]$ (Benjamini and Hochberg 1995; Lei and Fithian 2018). However, this assumption does not hold for the CRT p-values as they are in general not independent. They

are also in general only super-uniform under the null, but for all of the test statistics and settings considered in this paper the CRT's p-values are indeed exactly uniform under the null.

A key result is that under certain conditions, BH and AdaPT applied to the CRT p-values have the same asymptotic FDR and power as if the p-values were actually independent. Due to the cumbersome notation required, a formal presentation is deferred to Theorem 10 in Appendix E, where we give conditions on input p-values such that BH and AdaPT perform asymptotically as if the input p-values were independent.[2] Here, we only show the following Theorem 5 that applies Theorem 10 to the CRT with the three statistics considered in Section 2. In order to state Theorem 5, we need Definition 2, which allows us to concisely and intuitively characterize the asymptotic power expressions derived from Theorem 10. We note that in addition to characterizing the power, these two theorems are the first that we know of to prove asymptotic FDR control of multiple testing with CRT p-values.

Definition 2. *Let $\mathcal{P}$ be a multiple testing procedure that takes a set of p-values as input, e.g., the BH procedure at level q. Let $\mathcal{P}(\pi_\mu)$ be the procedure that applies $\mathcal{P}$ to p-values $\mathrm{pval}_1, \ldots, \mathrm{pval}_p$ in the following independent normal means model: for $j = 1, 2, \ldots, p$,*

$$\mu_j \overset{\text{i.i.d.}}{\sim} \pi_\mu, \qquad \varepsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \qquad \mathrm{pval}_j = 1-\Phi(\mu_j+\varepsilon_j) \quad (\text{respectively, } \mathrm{pval}_j = 2(1-\Phi(|\mu_j+\varepsilon_j|))).$$

We say a variable selection procedure has one-sided (respectively, two-sided) effective π_μ with respect to $\mathcal{P}$ if, as $p \to \infty$,

(a) the realized power (i.e., the proportion of rejected non-nulls) of this variable selection procedure converges in probability to the same constant that the realized power of $\mathcal{P}(\pi_\mu)$ converges in probability to; and

(b) when the asymptotic realized power is positive, the FDP of this variable selection procedure converges in probability to the same constant that the FDP of $\mathcal{P}(\pi_\mu)$ converges in probability to.

Theorem 5. *In Setting 2, for Lebesgue-almost-every $q \in (0,1)$, BH or AdaPT at level q using CRT p-values based on the statistics in Section 2.2 (respectively, their absolute values) have the following one-sided (respectively, two-sided) effective π_μ's with respect to BH or AdaPT at level q:*

1. *For the marginal covariance statistic, the effective π_μ is the distribution of $\frac{1}{\sqrt{\sigma^2+\kappa\mathbb{E}[B_0^2]}}B_0$.*
2. *For the OLS statistic, assuming $\kappa < 1$, the effective π_μ is the distribution of $\frac{\sqrt{1-\kappa}}{\sigma}B_0$.*
3. *For the distilled lasso statistic, the effective π_μ is the distribution of $\frac{1}{\tau_\lambda}B_0$.*

The effective π_μ's in Theorem 5 follow from Theorems 1, 2 and 3. The key component of the proof of Theorem 5 is jointly analyzing the test statistics for two different covariates showing that its two coordinates are asymptotically independent. This is particularly non-trivial for the distilled-lasso statistic, where we employ a leave-one-out approach. As one would expect, these procedures have higher power if the respective CRT with the same statistic has higher power. For example, for the marginal covariance statistic, as $\sigma^2 + \kappa\mathbb{E}[B_0^2]$ gets smaller, the null and non-null distributions of the p-values are more separated and higher power would be obtained. Naturally,

[2]Theorem 10 represents a variation on results of Ferreira et al. (2006) but with a different proof catered to our specific setting.

this is the same condition under which the CRT with the marginal covariance statistic has higher power, once we realize that $v_Z^2 = \kappa \mathbb{E}[B_0^2]$ (see the text immediately after Setting 1). Although we choose BH and AdaPT as representatives, we note that the same proof techniques could be used to establish analogous results for other procedures such as Storey's BH (Storey et al. 2004) that use the empirical distribution of the p-values in a certain way.

3.2 Model-X knockoffs

3.2.1 Review of knockoffs

We now turn to the analysis of model-X knockoffs (Candès et al. 2018), beginning with a review of the knockoffs procedure.

Consider again the regression setting where our data is composed of $[\mathbf{X}, \mathbf{Y}]$, whose rows are i.i.d. copies of $(X, Y) \sim F_{X,Y}$. The (eponymous) first step of the knockoffs procedure is to generate knockoffs. We say the $n \times p$ random matrix $\tilde{\mathbf{X}}$ is a knockoff matrix for $\mathbf{X}$ if $\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{X}$ and the following *pairwise exchangeability* is satisfied for each j:

$$[\mathbf{X}, \tilde{\mathbf{X}}] \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(j)},$$

where the subscript swap(j) denotes swapping the jth and $(j+p)$th columns of a matrix or elements of a vector (in this case, swapping $\mathbf{X}_j$ and $\tilde{\mathbf{X}}_j$). In Setting 2, because the covariates are independent, generating such knockoffs is particularly simple: we can just take $\tilde{\mathbf{X}}$ to be an i.i.d. copy of $\mathbf{X}$.

The second step is to define a variable importance statistic

$$T := T([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{Y}) = (T_1, \dots, T_p, \tilde{T}_1, \dots, \tilde{T}_p),$$

which satisfies

$$T([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(j)}, \mathbf{Y}) = (T_1, \dots, T_p, \tilde{T}_1, \dots, \tilde{T}_p)_{\text{swap}(j)}.$$

That is, swapping the column corresponding to the jth covariate X_j with that of its knockoff $\tilde{X}_j$ will swap their corresponding variable importance statistics T_j and $\tilde{T}_j$ and leave the other elements of T unchanged. A typical example of T is the absolute value of the fitted lasso coefficient vector from regressing $\mathbf{Y}$ on $[\mathbf{X}, \tilde{\mathbf{X}}]$. T is then plugged into an antisymmetric function $f(\cdot, \cdot)$ (i.e., $f(x, y) = -f(y, x)$) to compute $W \in \mathbb{R}^p$: $W_j = f(T_j, \tilde{T}_j)$. For example, we can simply let $f(x, y) = x - y$.

The third step is variable selection. It was shown in Candès et al. (2018) that if we select the set of variables

$$\hat{S} = \{j : W_j \geq \hat{w}\}, \quad \text{where} \qquad \hat{w} = \min\left\{w > 0 : \frac{1 + |\{j : W_j \leq -w\}|}{|\{j : W_j \geq w\}|} \leq q\right\},^3 \tag{2}$$

then the FDR is controlled at level q.

3.2.2 Marginal covariance and ordinary least squares variable importance statistics

A peculiarity of knockoffs is that its rejections are not determined by a vector of unordered p-values, but instead a ordered vector of signs, which could be viewed as "one-bit" p-values with an order. Thus, it is worthwhile to pause and discuss its relationship with p-values. As outlined in Section 3.2.1, knockoffs operates on unordered feature importance statistics $W_1, \dots, W_p$. If all the null W_j's have the same marginal distribution, let F be its CDF and consider oracle p-values given

[3] Formally, we make the minimum well-defined by only considering the minimum over $w \in \{|W_j| : W_j \neq 0, 1 \leq j \leq p\}$.

by $p_j = 1 - F(W_j)$ (such p-values cannot be computed in practice because F is unknown). Knockoffs with nominal FDR level q rejects all p-values below $\hat{t}_{\mathrm{KF}}$, where

$$\hat{t}_{\mathrm{KF}} = \max\left\{t \in [0, \frac{1}{2}) : \frac{1 + \#\{j : p_j \geq 1 - t\}}{\#\{j : p_j \leq t\}} \leq q\right\}.$$

This is equivalent to the intercept-only AdaPT procedure applied to the p_j's (Lei and Fithian 2018). Thus, we can understand the asymptotic behavior of the knockoffs procedure by studying the joint distribution of the p_j's. In fact, Theorem 11 in Appendix E shows that under certain conditions, we can treat the p_j's as independent draws from their respective asymptotic marginal distributions.

In proving the expressions for the asymptotic power of multiple testing with CRT p-values, we needed to analyze the asymptotic distributions of pairs of test statistics, and it turns out the same tools are sufficient for both establishing the assumptions of Theorem 11 and for characterizing the marginal distributions of the p_j's, except that analysis of the asymptotic distributions of sets of *four* test statistics is needed. In particular, Lemma 13 in Appendix E says that we just need to check that for distinct j and k, $(T_j, T_k, \tilde{T}_j, \tilde{T}_k)$ converges in distribution to a random vector with independent coordinates in order for Theorem 11 to hold.

While our asymptotic analysis of $(T_j, T_k, \tilde{T}_j, \tilde{T}_k)$ for a given statistic allows us to obtain the asymptotic power of knockoffs for any antisymmetric function f, when $f(x, y) = x - y$, if the test statistic is the marginal covariance or the OLS coefficient, there is a direct and easily interpretable connection to the AdaPT procedure applied to a normal means model, and we can state our results using the language of effective π_μ from Definition 2.

Theorem 6. *In Setting 2, for almost every $q \in (0, 1)$, knockoffs with $\tilde{\mathbf{X}}$ an i.i.d. copy of $\mathbf{X}$ and the antisymmetric function $f(x, y) = x - y$ at level q with marginal covariance or OLS test statistic has the following one-sided effective π_μ's with respect to the AdaPT procedure at level q:*

1. *For the marginal covariance statistic, the effective π_μ is the distribution of $\frac{1}{\sqrt{2(\sigma^2 + \kappa\mathbb{E}[B_0^2])}} B_0$.*
2. *For the OLS statistic, assuming $\kappa < 1/2$, the effective π_μ is the distribution of $\frac{\sqrt{1-2\kappa}}{\sqrt{2\sigma^2}} B_0$.*

Note that our general result Theorem 11 covers the two-sided case, which is equivalent to taking $f(x, y) = |x| - |y|$, but it cannot be expressed in terms of an effective π_μ. We can observe that the effective π_μ's in Theorem 6 agree with Theorems 1 and 2. We see that knockoffs with the OLS statistic outperforms knockoffs with the marginal covariance statistic if $\sigma^2/(1 - 2\kappa) < \sigma^2 + \kappa\mathbb{E}[B_0^2]$, and vice versa. Comparing Theorems 6 and 5, we can see that multiple testing with CRT p-values effectively increases the signal size by a factor of $\sqrt{2}$ compared to knockoffs for the marginal covariance. For OLS, multiple testing with CRT p-values effectively increases the signal size by a factor of $\sqrt{2}\sqrt{(1-\kappa)/(1-2\kappa)} > \sqrt{2}$ over knockoffs, with the additional factor $\sqrt{(1-\kappa)/(1-2\kappa)}$ approaching infinity as $\kappa \to 1/2$ from below.

3.2.3 Knockoffs with the lasso coefficient

The lasso coefficient is a popular statistic frequently used with knockoffs. Specifically, let $\hat{\beta}^\lambda$ be the coefficient estimate from using the lasso to regress $\mathbf{Y}$ on $[\mathbf{X}, \tilde{\mathbf{X}}]$ with penalty parameter λ. We suppress the superscript λ when there is no confusion. For $j = 1, 2, \ldots, p$, let $T_j = \sqrt{n}\hat{\beta}_j$, $\tilde{T}_j = \sqrt{n}\hat{\beta}_{j+p}$, and $W_j = f(T_j, \tilde{T}_j)$ for some antisymmetric function f.

Theorem 7. *In Setting 2, knockoffs with* $\tilde{\mathbf{X}}$ *an i.i.d. copy of* $\mathbf{X}$*, antisymmetric function* $f(x, y)$ *satisfying the mild regularity condition in Theorem 12, and variable importance statistic* $\hat{\beta}^{\lambda}$*, at Lebesgue-almost-every level* $q \in (0, 1)$*, has the same asymptotic power as if the* $\sqrt{n}\hat{\beta}^{\lambda}_j$*'s were independent, where for* $j = 1, 2, \ldots, p$*,* $\sqrt{n}\hat{\beta}^{\lambda}_j \sim \eta(B_0 + \tau_\lambda Z; \alpha_\lambda \tau_\lambda)$ *and* $\sqrt{n}\hat{\beta}^{\lambda}_{j+p} \sim \eta(\tau_\lambda Z; \alpha_\lambda \tau_\lambda)$*,* $Z \sim \mathcal{N}(0, 1)$ *independent of* B_0*.*

See Theorem 12 in Appendix E for a detailed presentation. We prove the asymptotic independence via a symmetry argument, and extra care is taken due to the fact that the $\hat{\beta}_j$'s have a delta mass at 0. During the preparation of this manuscript, we discovered an independent and parallel work on the asymptotic power of knockoffs using the lasso coefficient difference statistic (Weinstein et al. 2020), which provides a nearly identical power result to our Theorem 7.

Now we heuristically compare the asymptotic power of knockoffs with the lasso coefficient with that of multiple testing with CRT p-values obtained from the distilled lasso statistic. Since the two results involve two different τ_λ's, we differentiate them with $\tau^{\mathrm{CRT}}_{\lambda_{\mathrm{CRT}}}$ and $\tau^{\mathrm{KF}}_{\lambda_{\mathrm{KF}}}$, respectively (note that the two are generally different even when $\lambda_{\mathrm{CRT}} = \lambda_{\mathrm{KF}}$, as they also implicitly depend on other parameters). From Theorem 7, we can interpret τ_λ as the standard deviation of the noise added to the signal B_0, with a thresholding operation afterwards. On the other hand, we see from Theorem 5, by a rescaling of mean and variance, $\tau^{\mathrm{CRT}}_{\lambda_{\mathrm{CRT}}}$ as the standard deviation of noise added to the signal B_0. It turns out that if we choose the best oracle λ for the CRT, $\tau^{\mathrm{CRT}}_{\lambda_{\mathrm{CRT}}} \leq \tau^{\mathrm{KF}}_{\lambda_{\mathrm{KF}}}$, because knockoffs doubles the dimension of covariates and thus introduces more noise (see Appendix H for a formal proof). Intuitively, we should expect higher power from the CRT. We provide a numerical comparison in Section 3.3.

3.3 Asymptotic power comparison of multiple testing with CRT p-values and knockoffs

When the marginal covariance or the OLS coefficient is used as the variable importance statistic and the antisymmetric function is $f(x, y) = x - y$, Theorem 6 provides a direct comparison between the asymptotic power of knockoffs with that of multiple testing with CRT p-values (see the text after Theorem 6). In practice, we usually do not know the signs of the signals, so it is more common to use the absolute value of the marginal covariance, the OLS coefficient, or the lasso coefficient as the variable importance statistic (or, equivalently, take $f(x, y) = |x| - |y|$). These results do not fit into Definition 2 with an effective π_μ because asymptotically, although the test statistics are independent, they are not marginally Gaussian. In this section, we numerically compare these results with the power of multiple testing with CRT p-values. We see in Figures 2, 3, and 4 that knockoffs is less powerful than the CRT methods. It is interesting that in lower-dimensional settings such as $n = 2.5p$, knockoffs with two-sided test statistics is more powerful than the $\sqrt{2}$-signal strength reduction relative to the CRT suggested by the analysis for one-sided test statistics in Section 3.2.2, and for the lasso statistic, there is almost no power difference in such lower-dimensional settings.

4 Retrospective sampling

As a generalization of our results for the CRT in Section 2, we consider a case in which we know the distribution of $\mathcal{L}(X \mid Z)$, but the data have been collected retrospectively. Specifically, we assume the following model.

Setting 3 (High-dimensional linear model with retrospective sampling)**.** *Let* $g : \mathbb{R} \to [0, 1]$ *be a Borel function that is not almost everywhere* 0*. For each* p*, generate i.i.d. data from Setting 1 and*

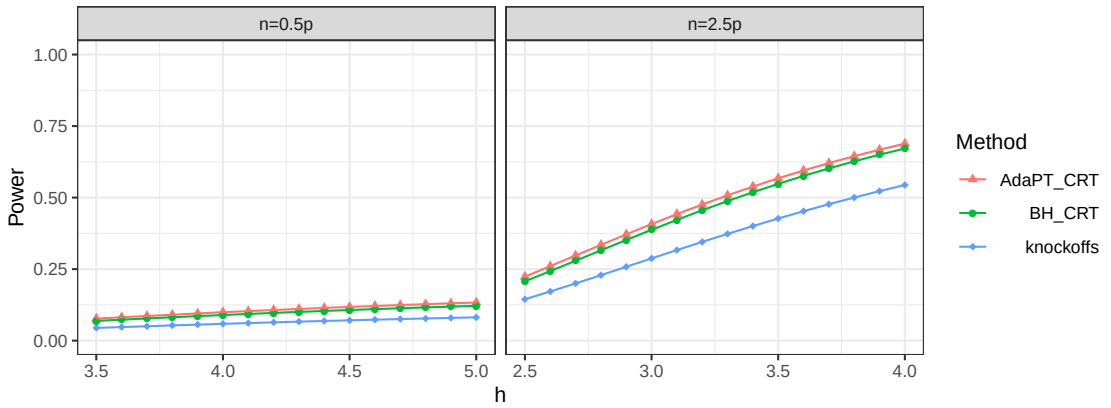

Figure 2: Asymptotic power comparison for BH and AdaPT applied to two-sided CRT p-values and knockoffs with the absolute value of the marginal covariance with the original signal size and $\sqrt{2}$ times the signal size. Plot is in Setting 2 with $\gamma = 0.9$ and $\pi_1 = \delta_h$ with varying h, and the nominal FDR level is 0.1.

reject each data point (X_i, Y_i, Z_i) with probability $1 - g(Y_i)$, until n data points have been collected, such that $p/n \to \kappa$ still holds.

Barber and Candès (2019, Proposition 1) established that the CRT remains valid when Setting 1 is assumed but the data actually come from Setting 3. In addition to single hypothesis testing in Setting 3, we will also consider the variable selection with the CRT p-values, coming from Setting 4 as follows.

Setting 4 (High-dimensional linear model with retrospective sampling)**.** *Let $g : \mathbb{R} \to [0, 1]$ be a Borel function that is not almost everywhere* 0*. For each p, generate i.i.d. data from Setting 2 and reject each data point (X_i, Y_i) with probability $1 - g(Y_i)$, until we have n data points, such that $p/n \to \kappa$ still holds.*

Barber and Candès (2019) also established that the knockoffs are still valid when Setting 2 is assumed but the data actually come from Setting 4. Thus, in this section, we will consider knockoffs generated independently as in Section 3.2. The following theorem gives the asymptotic power of the CRT and knockoffs using the marginal covariance statistic with retrospective sampling.

Theorem 8. *Consider using the test statistics $T = n^{-1}\mathbf{X}^\top\mathbf{Y}$ for the CRT, and $T_j = n^{-1}\mathbf{X}_j^\top\mathbf{Y}$ for multiple testing with CRT p-values and knockoffs. Let M^2_{retro} be the asymptotic second moment of the retrospectively collected Y_i, i.e.,*

$$M^2_{\text{retro}} = \frac{\mathbb{E}[Y^2_{\text{raw}} g(Y_{\text{raw}})]}{\mathbb{E}[g(Y_{\text{raw}})]},$$

where $Y_{\text{raw}} \sim \mathcal{N}(0, \sigma^2 + v_Z^2)$ is drawn from the asymptotic distribution of Y without rejection.[4] *Note that in Setting 4, the corresponding v_Z^2 (or $v_{X_{\text{-}j}}^2$) is equal to $\kappa\mathbb{E}[B_0^2]$.*

1. *In Setting 3, the asymptotic power of the CRT is equal to that of a z-test with standardized effect size*

$$\frac{hM_{\text{retro}}}{v_Z^2 + \sigma^2}.$$

[4] M_{retro} always exists because $g(y) \in [0, 1]$ and is not almost everywhere zero.

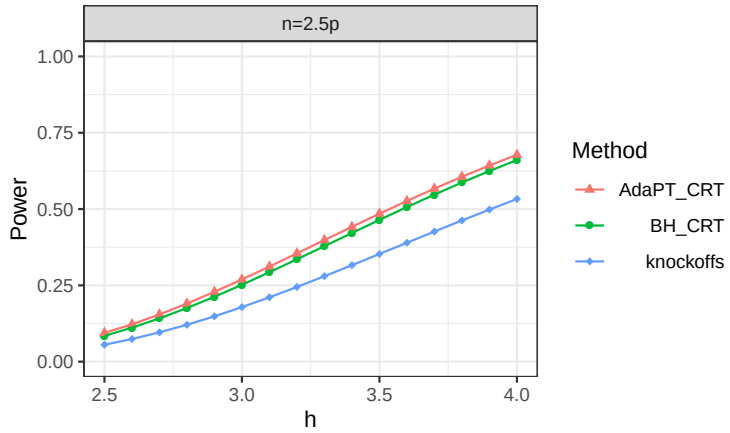

Figure 3: Asymptotic power comparison for BH and AdaPT applied to two-sided CRT p-values and knockoffs with the absolute value of the OLS coefficient with the original signal size and $\sqrt{2}$ times the signal size. The setting is the same as in Figure 2.

2. *In Setting 4, for almost all $q \in (0,1)$, BH or AdaPT at level q applied to CRT p-values using T_j (or $|T_j|$) have one-sided (or two-sided) effective π_μ given by the distribution of $\frac{M_{\text{retro}}}{\sigma^2+\kappa\mathbb{E}[B_0^2]}B_0$ with respect to BH or AdaPT at level q.*

3. *In Setting 4, for almost all $q \in (0,1)$, knockoffs with $\tilde{\mathbf{X}}$ an i.i.d. copy of $\mathbf{X}$, antisymmetric function $f(x,y) = x - y$, test statistic T_j, and level q has one-sided effective π_μ given by the distribution of $\frac{M_{\text{retro}}}{\sqrt{2}(\sigma^2+\kappa\mathbb{E}[B_0^2])}B_0$ with respect to AdaPT at level q.*

To sum up, Theorem 8 establishes that for retrospective sampling, the same results on the asymptotic power hold with the signal size multiplied by $\frac{M_{\text{retro}}}{\sqrt{\sigma^2+v_Z^2}}$. Thus, the power gets higher as M_{retro} gets larger. This is intuitive, since it should be easier for us to detect the signal in regions where Y has extreme values. As a special case, if $g \equiv 1$, then $M_{\text{retro}} = \sqrt{\sigma^2 + v_Z^2}$ and we return to the non-retrospective sampling case.

While the asymptotic power expressions for retrospective sampling can be higher than that of non-retrospective sampling, it comes at a price of requiring more raw samples, and it is worthwhile to discuss the implications. Let n_{raw} be the number of raw samples needed to get n retrospective samples, then $n/n_{\text{raw}} \to \int \phi_{\sigma^2+v_Z^2}(y)g(y)\,dy$. If we do not discard any samples and use all n_{raw}, we return to the non-retrospective sampling settings with h increased to $h/\sqrt{\int \phi_{\sigma^2+v_Z^2}(y)g(y)\,dy}$ (or B_0 to $B_0/\sqrt{\int \phi_{\sigma^2+v_Z^2}(y)g(y)\,dy}$). One can then directly compare the asymptotic powers and note that, as intuition would suggest, the power is maximized when no sample is rejected. In practice, however, collecting covariates might be expensive. Therefore, it can be beneficial to decide whether or not to collect the covariates based on a screening step using the value of Y. A natural question is then how to achieve the highest power while fixing the sampling cost. This is equivalent to maximizing M_{retro} while fixing $\int \phi_{\sigma^2+v_Z^2}(y)g(y)\,dy$ and it is not hard to see that the maximum is attained when $g(y) = \mathbf{1}_{|y|>C}$ for an appropriate C.

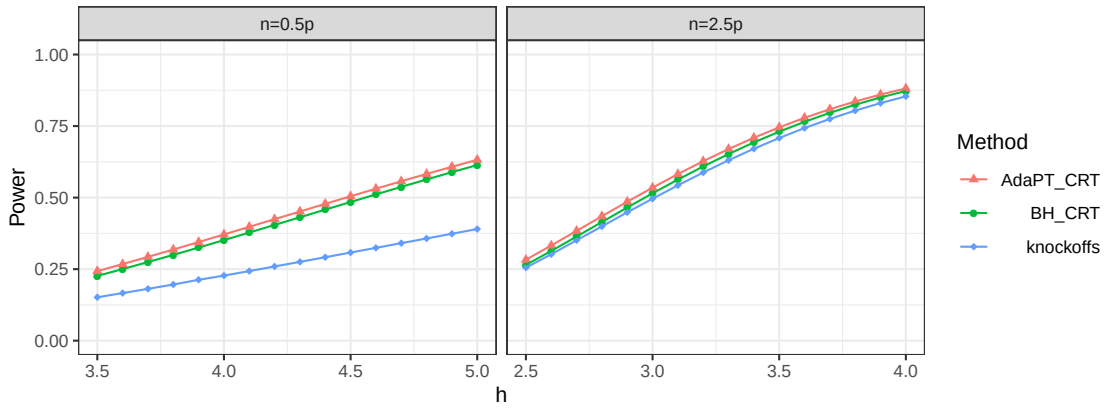

Figure 4: Asymptotic power comparison for BH and AdaPT applied to two-sided CRT p-values with the distilled lasso statistic and knockoffs with the absolute value of the lasso coefficient with the original signal size and $\sqrt{2}$ times the signal size. The setting is the same as in Figure 2. For all methods, λ is chosen so that the asymptotic power is maximized (for the CRT, this is equivalent to minimizing τ_λ).

5 Simulations

In this section, we examine the finite-sample accuracy of our asymptotic power expressions.

5.1 CRT in Setting 1

In Figure 5, we compare the power of the CRT with each statistic mentioned in Section 3.1. We plot as a horizontal line the power of the CRT with an oracle statistic that is the upper bound for the achievable power with the CRT (see Appendix I.1.2). We can see that the distilled lasso statistic has comparable power with the optimal statistic.

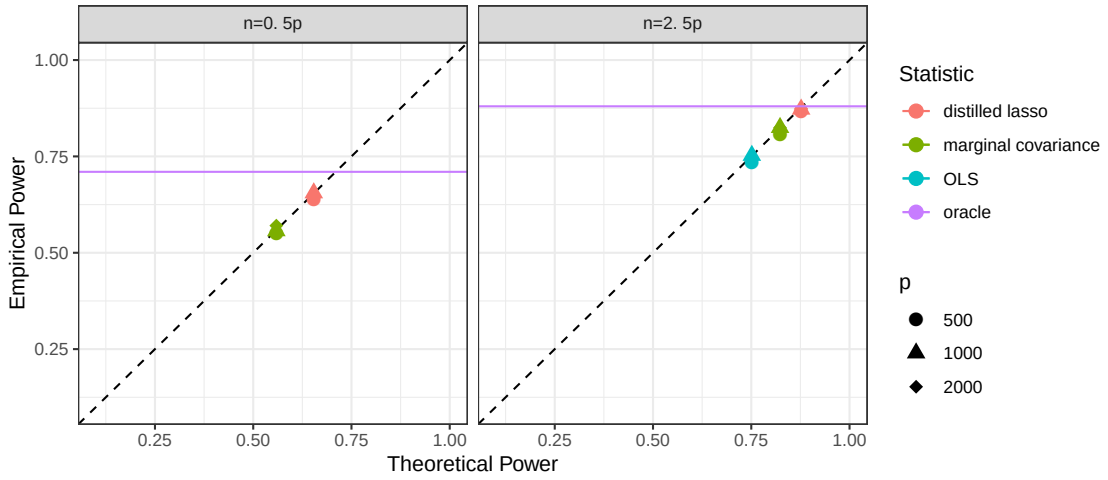

Figure 5: Comparison of the powers of the CRT using different statistics. The setting is Setting 1 with $\sigma^2 = 1$, $\xi = 0$, $\beta = 3$, $\sqrt{n}\theta_j \overset{\text{i.i.d.}}{\sim} 0.9\delta_0 + 0.1\delta_3$. The results for Bayes are empirical based on 960 independent simulations. For the distilled lasso statistic, λ is chosen so that τ_λ is minimized for highest asymptotic power. All standard errors are below 0.01.

5.2 Conjecture 1

In this section, we show simulation results regarding Conjecture 1 in Section 2.3. In Figure 6, we plot the conjectured asymptotic power and empirical finite-sample power as a function of n_*/p with $n/p = 1.5$ fixed for two different values of v_Z^2. Note that the conjectured power must agree with the empirical power in the limit as p and n_*/p go to infinity (Theorem 4).

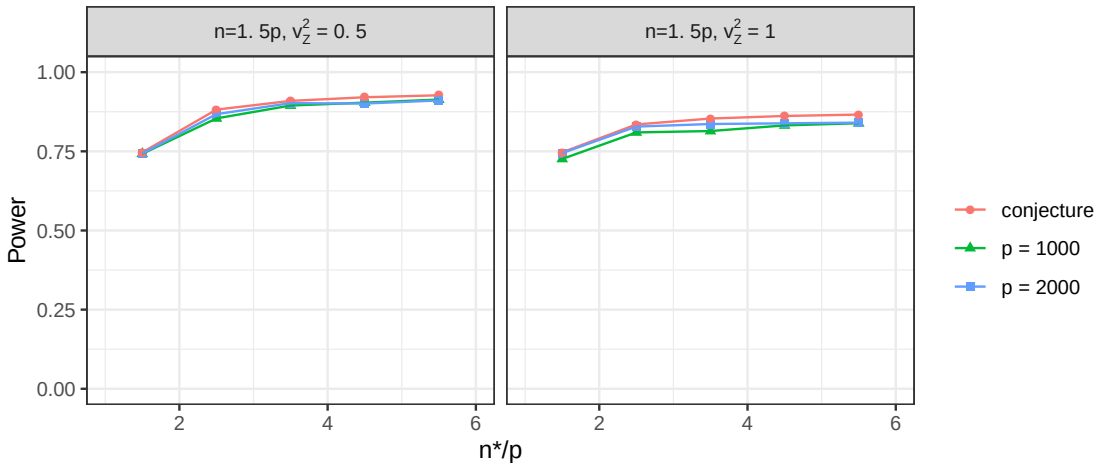

Figure 6: Simulations are for $p = 1,000$ and $p = 2,000$ with $h = 4$ in the setting of Conjecture 1. All standard errors are below 0.01.

5.3 Multiple testing with CRT p-values and knockoffs

In this section, we show some simulation results of BH applied to CRT p-values (BH-CRT) and knockoffs with the statistics discussed in this paper. We defer the results of AdaPT applied to CRT p-values to Appendix I.2 so we do not crowd the plots; in summary, AdaPT has slightly higher power and FDR and converges more slowly than BH, especially in low-power settings. In our simulations, $\gamma = 0.9$ and π_1 is a point mass at $h = 4$. We use absolute values of the statistics, since in practice we do not know the sign of h. Points with the same color represent methods with the same statistic and different p's, including $p = \infty$, which is calculated based on our theory. It can be seen that points of the same color form separated clusters, which means that our theory could guide statistic choices even in finite samples. We note that the finite-sample agreement is not quite as good for knockoffs in lower-power settings as that for the BH-CRT, because of the discreteness in the numerator of the FDP estimate in the knockoffs procedure (the fraction in Equation (2)). We also include the results of an oracle using the Bayesian method that controls the Bayesian FDR (see Appendix I.1.1), and BH-CRT with the distilled lasso statistic can be close to this oracle method when $n = 2.5p$, while when $n < p$, there is still a substantial gap as would be expected given the relative value of the prior to the smaller sample size.

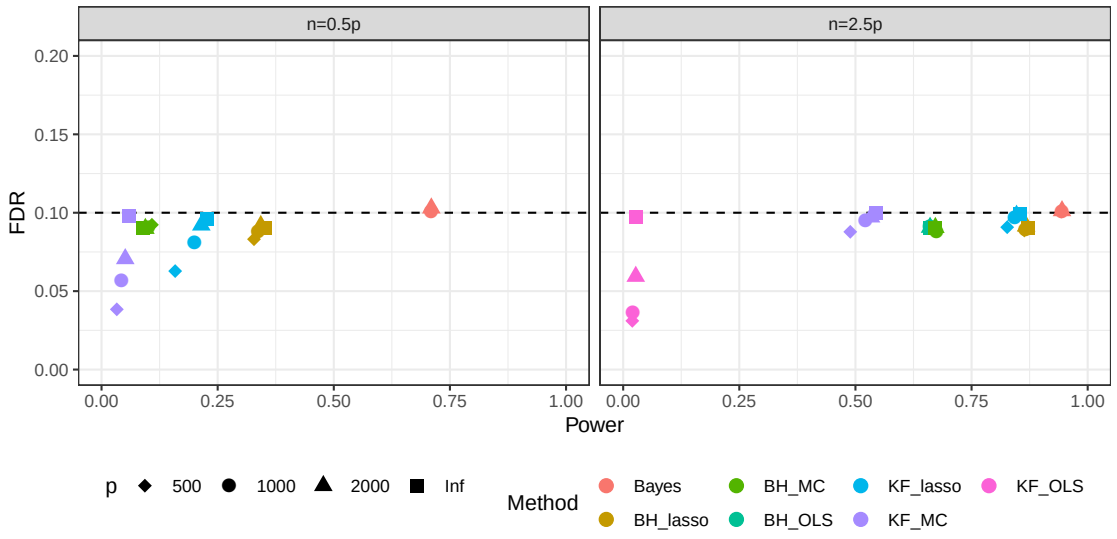

Figure 7: Power comparison of different methods at FDR level 0.1 with the same setting as that of Figure 4. For the lasso statistics, λ is chosen so that the asymptotic power is maximized (for the CRT, this is equivalent to minimizing τ_λ), while we note that the specific choice of λ only affects the power mildly within a reasonable range (see Figure 1). All standard errors are below 0.01.

5.4 Retrospective sampling

In this section, we compare the empirical and theoretical powers of the CRT in Setting 3, where g is taken to be of the form $g(x) = \mathbf{1}_{|x|>\text{threshold}}$ for different thresholds.

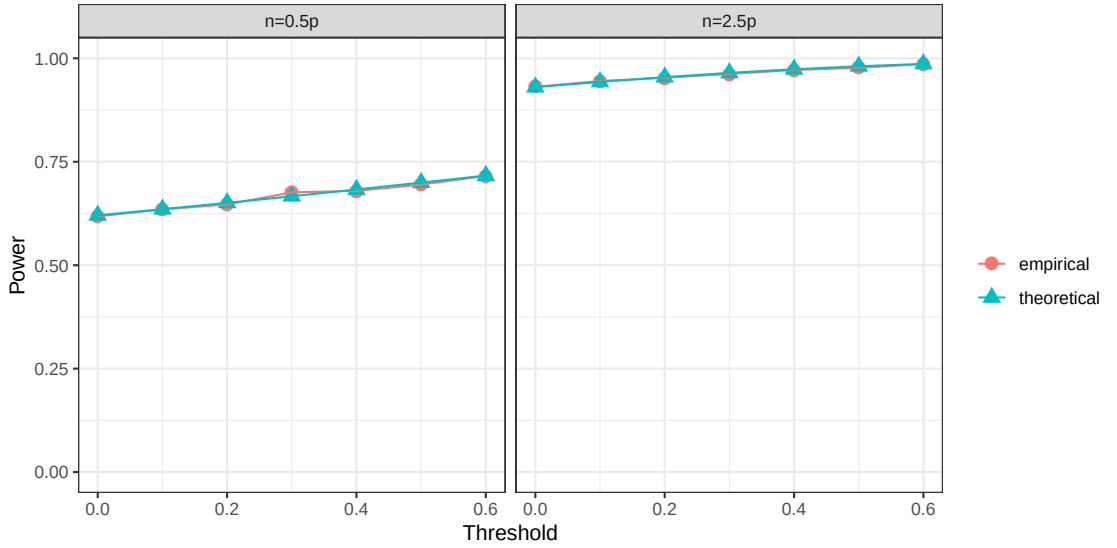

Figure 8: Comparison of the empirical ($p = 500$) and theoretical (asymptotic) powers of the CRT with the marginal covariance statistic in the retrospective sampling setting (Setting 3 with $\sigma^2 = 1$, $\xi = 0$, $\beta = 4$ and $\sqrt{n}\theta_j \overset{\text{i.i.d.}}{\sim} 0.9\delta_0 + 0.1\delta_4$). All standard errors are below 0.01.

6 Discussion

This paper studied the asymptotic powers of the CRT and knockoffs in the high-dimensional regime, i.e., as $n, p \to \infty$, n/p goes to a positive constant and a fixed non-zero proportion of variables are non-null. A very natural future direction is to study the behavior of the CRT and knockoffs with different statistics and/or in other settings. For example, Celentano et al. (2020) could provide starting points on extending our lasso power analysis to settings with correlated covariates, while Sur and Candès (2019); Liang and Sur (2020) could enable the study of binary regression settings and their corresponding test statistics. Alternatively, the power analysis of oracle test statistics (e.g., the one in Appendix I.1.2) could provide theoretical bounds on the power of these methods with any statistics.

Acknowledgements

The authors would like to thank Hong Hu, Tracy Ke, Natesh Pillai, Subhabrata Sen, and Pragya Sur for valuable discussions and suggestions. L. J. was partially supported by the William F. Milton Fund.

References

Barber, R. F. and Candès, E. (2019). On the construction of knockoffs in case–control studies. *Stat*, 8(1):e225.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2018). Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*.

Bates, S., Candès, E. J., Janson, L., and Wang, W. (2020a). Metropolized knockoff sampling. *Journal of the American Statistical Association*.

Bates, S., Sesia, M., Sabatti, C., and Candès, E. (2020b). Causal inference in genetic trio studies. *arXiv preprint arXiv:2002.09644*.

Bayati, M. and Montanari, A. (2011). The LASSO risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, pages 289–300.

Berrett, T. B., Wang, Y., Barber, R. F., and Samworth, R. J. (2019). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: Model-X knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B*, 80(3):551–577.

Celentano, M., Montanari, A., and Wei, Y. (2020). The lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chia, C., Sesia, M., Ho, C.-S., Jeffrey, S. S., Dionne, J., Candès, E. J., and Howe, R. T. (2020). Interpretable signal analysis with knockoffs enhances classification of bacterial raman spectra. *arXiv preprint arXiv:2006.04937*.

Cochran, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 30, pages 178–191. Cambridge University Press.

Fan, Y., Lv, J., Sharifvaghefi, M., and Uematsu, Y. (2018). IPAD: stable interpretable forecasting with knockoffs inference. *Available at SSRN 3245137*.

Ferreira, J., Zwinderman, A., et al. (2006). On the Benjamini–Hochberg method. *The Annals of Statistics*, 34(4):1827–1849.

Huang, D. and Janson, L. (2020). Relaxing the assumptions of knockoffs by conditioning. *The Annals of Statistics*.

Javanmard, A., Montanari, A., et al. (2018). Debiasing the lasso: Optimal sample size for Gaussian designs. *The Annals of Statistics*, 46(6A):2593–2622.

Katsevich, E. and Ramdas, A. (2020). A theoretical treatment of conditional independence testing under model-x. *arXiv preprint arXiv:2005.05506*.

Katsevich, E. and Roeder, K. (2020). Conditional resampling improves sensitivity and specificity of single cell crispr regulatory screens. *bioRxiv*.

Katsevich, E. and Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The annals of applied statistics*, 13(1):1.

Lehmann, E. L. and Romano, J. P. (2006). *Testing statistical hypotheses*. Springer Science & Business Media.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(4):649–679.

Li, J. and Maathuis, M. H. (2019). Nodewise knockoffs: False discovery rate control for gaussian graphical models. *arXiv preprint arXiv:1908.11611*.

Liang, T. and Sur, P. (2020). A precise high-dimensional asymptotic theory for boosting and min-ℓ_1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*.

Liu, J. and Rigollet, P. (2019). Power analysis of knockoff filters for correlated designs. In *Advances in Neural Information Processing Systems*, pages 15420–15429.

Liu, M., Katsevich, E., Janson, L., and Ramdas, A. (2020). Fast and powerful conditional randomization testing via distillation. *arXiv preprint arXiv:2006.03980*.

Lu, Y., Fan, Y., Lv, J., and Noble, W. S. (2018). DeepPINK: reproducible feature selection in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 8689–8699.

McMurdie, P. J. and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531.

Rencher, A. C. and Schaalje, G. B. (2008). *Linear models in statistics*. John Wiley & Sons.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, 31(4):567–584.

Sard, A. (1942). The measure of the critical values of differentiable maps. *Bulletin of the American Mathematical Society*, 48(12):883–890.

Sesia, M., Bates, S., Candès, E., Marchini, J., and Sabatti, C. (2020a). Controlling the false discovery rate in gwas with population structure. *bioRxiv*.

Sesia, M., Katsevich, E., Bates, S., Candès, E., and Sabatti, C. (2020b). Multi-resolution localization of causal variants across the genome. *Nature communications*, 11(1):1–10.

Sesia, M., Sabatti, C., and Candès, E. J. (2018). Gene hunting with hidden Markov model knockoffs. *Biometrika*.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Sur, P. and Candès, E. J. (2019). A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525.

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., and Blei, D. M. (2018). The holdout randomization test: Principled and easy black box feature selection. *arXiv preprint arXiv:1811.00645*.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Weinstein, A., Barber, R., and Candes, E. (2017). A power and prediction analysis for knockoffs with lasso statistics. *arXiv preprint arXiv:1712.06465*.

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.

Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: a tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 34(3):275–285.

Zhu, Y. and Bradic, J. (2018). Significance testing in non-sparse high-dimensional linear models. *Electronic Journal of Statistics*, 12(2):3312–3364.

A Notation

Bold letters are used for matrices or vectors containing i.i.d. observations. Unless specified otherwise, a vector is always a column vector instead of row vector. For a vector a, a_S denotes the sub-vector that consists of elements indexed by S; for a matrix A, $A_{S,S}$ denotes the sub-matrix that consists of rows and columns indexed by S. For integers $i \leq j$, the notation $i:j$ means the set $\{i, i+1, \ldots, j\}$, and we use $[p]$ to denote $1:p$. For a set $S \subseteq [p]$, $|S|$ denotes the number of elements in S, -S denotes the set $[p] \setminus S$. Let I_d be the $d \times d$ identity matrix and for $d_1 \leq d_2$, let $I_{d_1 \times d_2}$ be the matrix obtained by adding $(d_2 - d_1)$ rows of zeros to I_{d_1}. Let $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d, \|x\|_2 = 1\}$. The indicator function of B is denoted as $\mathbf{1}_B$, i.e., it takes value 1 on B and zero otherwise. The cumulative distribution function (CDF) of the Gaussian distribution $\mathcal{N}(0,1)$ is denoted by Φ—for $\alpha \in (0,1)$, z_α denotes the α-quantile of $\mathcal{N}(0,1)$, i.e., $\Phi(z_\alpha) = \alpha$. We use χ^2_k and Inv-χ^2_k to denote the chi-squared distribution and inverse chi-squared distribution with k degrees of freedom. For random variables or vectors W_1 and W_2, $\mathcal{L}(W_1)$ means the distribution of W_1 and $\mathcal{L}(W_1 \,|\, W_2)$ means the conditional distribution of W_1 given W_2. To ease notation when analyzing the power and false discovery rate, we use the convention that $0/0$ is defined to be 0. Unless another measure is explicitly specified, "almost everywhere" or "almost every" is with respect to the Lebesgue measure.

B CRT under low-dimensional asymptotics

As a side note, we consider a case in which we test a scalar parameter with no nuisance parameters under the asymptotics of local alternatives. One can think of this case as testing if a coefficient is zero in a linear regression setting, where the other coefficients are known. A similar problem was studied in Katsevich and Ramdas (2020), the difference of which we will discuss below.

We consider the setting with i.i.d. data $(X_i, Y_i, Z_i)_{i=1}^n = (\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Recall that (X, Y, Z) is actually simplified notation for (X_j, Y, X_{-j}). The null distribution is $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \sim P^n_{\theta_0}$, where $X \perp\!\!\!\perp Y \mid Z$ under P_{θ_0}. The alternative distribution is $(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \sim P^n_{\theta_0 + hn^{-1/2}}$, where h is a fixed scalar. We assume P_θ is q.m.d. and thus the two sequences are contiguous (see Appendix D.1). In other words, we are testing $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_0 + h/\sqrt{n}$ with n independent draws from P_θ. For presentational simplicity, suppose we know the sign of h is positive, while the case where we do not know the sign of h can be similarly studied. We remark that although contiguity gives us an interesting setting to analyze non-trivial power, it is not a necessary condition (see Appendix D).

Asymptotically linear statistics are an important class of statistics, which are of the form

$$T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = n^{-1/2} \sum_{i=1}^n \psi(X_i, Y_i, Z_i) + o_{\mathbb{P}_{H_0}}(1),$$

where $o_{\mathbb{P}_{H_0}}(1)$ denotes a term that goes to zero in probability under H_0. Many statistics can be written in this form, e.g., the log-likelihood ratio statistic and the score statistic. We will see that this class of statistics also can offer a most powerful test.

As suggested in Section 2.1, the CRT is run by finding a cutoff $c_\alpha(\mathbf{Y}, \mathbf{Z})$ such that we get an exactly size-α test conditional on $(\mathbf{Y}, \mathbf{Z})$ by rejecting when $T_n < c_\alpha$, accepting when $T_n > c_\alpha$, and randomizing when $T_n = c_\alpha$.

Theorem 9. *The asymptotic unconditional power of the above test under the local alternatives is*

$$1 - \Phi\left(z_{1-\alpha} - h\sqrt{\mathrm{Var}_{H_0}(s(X_i, Y_i, Z_i, \theta_0))}\,\mathrm{Corr}_{H_0}\left(\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i), s(X_i, Y_i, Z_i, \theta_0)\right)\right),$$

where,

$$e_0(Y, Z) = \mathbb{E}_{H_0}[\psi(X, Y, Z) \,|\, Y, Z]$$

and s is the score function, which, under very general regularity conditions,[5] *admits the common form*

$$s(X, Y, Z, \theta_0) = \frac{\frac{\partial}{\partial\theta}\big|_{\theta=\theta_0} p_\theta(X, Y, Z)}{p_{\theta_0}(X, Y, Z)},$$

where p_θ is the density of P_θ.

Let $\varphi_\psi(X_i, Y_i, Z_i) = \psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i)$. We see that to achieve high power, we need to find a ψ such that φ_ψ is highly correlated with $s(X_i, Y_i, Z_i, \theta_0)$.

Remark 1. If $\mathbb{E}_{H_0}[s(X, Y, Z, \theta_0) \,|\, Y, Z] = 0$, which is satisfied when the distribution of (Y, Z) does not depend on θ (but this is not necessary), then we can use $\psi = s$ itself and achieve the optimal asymptotic power (this is also the Neyman–Pearson statistic and achieves the unconditional optimal asymptotic power; see Example 12.3.12 in Lehmann and Romano (2006)). This means the family of asymptotically linear statistics includes an asymptotically most powerful test if $\mathbb{E}_{H_0}[s(X, Y, Z, \theta_0) \,|\, Y, Z] = 0$. This partially answers the question about model-X optimality in Remark 1 of Katsevich and Ramdas (2020) (i.e., the CRT with the score statistic is optimal among all valid tests in a certain asymptotic regime), which can also be seen as a generalization of the discussion "A precise parallel with OLS" in their Section 5.3 to non-linear regression settings.

Remark 2. Notation-wise, X and Y are symmetric, and thus the same result holds if we swap X and Y, which actually corresponds to the traditional fixed-X test, i.e., a test that is valid conditional on the covariates (X, Z). Since $\mathbb{E}_{H_0}[s(X, Y, Z, \theta_0) \,|\, X, Z] = 0$ always holds when (X, Z) is the covariate and $\mathcal{L}(X, Z)$ does not depend on θ, we can always use $\psi = s$ to achieve the optimal asymptotic power in the fixed-X framework.

Remark 3. Consider using the maximum likelihood estimate (MLE)

$$\hat{\theta}_n = \arg\max_\theta p_\theta(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$$

as the test statistic, which is equivalent to using $\sqrt{n}(\hat{\theta}_n - \theta_0)$ that satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{-\frac{1}{\sqrt{n}} \sum_{i=1}^n \ell'(\theta_0 \,|\, X_i, Y_i, Z_i)}{\frac{1}{n} \sum_{i=1}^n \ell''(\theta' \,|\, X_i, Y_i, Z_i)}$$

for some θ' between θ_0 and $\hat{\theta}_n$, where ℓ is the log-likelihood function. Since $-n^{-1} \sum_{i=1}^n \ell''(\theta' \,|\, X_i, Y_i, Z_i)$ converges in probability to the Fisher information $I(\theta_0)$ under the null (thus also under the alternative by contiguity; see, e.g., Theorem 12.3.2 in Lehmann and Romano (2006)), we see that the standardized MLE is asymptotically equivalent to the score statistic up to a multiplicative constant. This signifies that it also enjoys the optimal asymptotic power under the same condition $\mathbb{E}_{H_0}[s(X, Y, Z, \theta_0) \,|\, Y, Z] = 0$.

Remark 4. This result is closely related to Theorem 1 in Katsevich and Ramdas (2020), and we would like to highlight the key differences. (a) Our result applies to a general distribution P_θ and a general asymptotic linear statistic ψ, while Katsevich and Ramdas (2020) assume $\mathcal{L}(Y \,|\, X, Z)$ is Gaussian and considers a family of score-like statistics. (b) We assume there is no nuisance parameter, which corresponds to knowing the function g in Katsevich and Ramdas (2020); there, a deterministic estimate $\hat{g}$ is used instead, and the accuracy of $\hat{g}$ explicitly affects the power.

[5] See Theorem 12.2.1 in Lehmann and Romano (2006) for an example of such conditions. There, the notation $\tilde{\eta}$ is used instead of s.

We wish to emphasize that it is not true that the fixed-X framework always provides an optimal test, as seemingly suggested by Remarks 1 and 2. Specifically, Appendix C.1 exhibits a case where no fixed-X test can have nontrivial power, while a model-X test can, and Appendix C.2 shows that when testing a scalar parameter without nuisance parameters in non-asymptotic regimes, the optimal test can be a model-X one instead of a fixed-X one.

C Simple examples

C.1 Example where fixed-X has no power

Despite the fact that the fixed-X framework has been more heavily studied, it is not always "better" than the model-X framework. In fact, we provide a simple toy example where model-X methods have to be used for non-trivial inference. Consider the regression model

$$\mathbf{Y} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{X}^\top \beta, I_n), X_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), i = 1, 2, \ldots, n, j = 1, 2, \ldots, p, n < p - 1.$$

Here, we use $\mathbf{X}$ to denote the $n \times p$ data matrix and $\mathbf{Y}$ to denote the $n \times 1$ response vector. Now suppose we would like to construct a fixed-X statistical test for $H_0 : \beta_1 = 0$. We claim that such a test must have trivial power. Formally, let $T_{\mathbf{X}}(\mathbf{Y})$ be a valid level-α test, i.e.,

$$\mathbb{P}(T_{\mathbf{X}}(\mathbf{Y}) = 1 \mid \mathbf{X}, \beta) \leq \alpha, \forall \beta \in \mathbb{R}^p \text{ s.t. } \beta_1 = 0. \tag{3}$$

To analyze its power, consider any γ where $\gamma_1 \neq 0$. There exists $\tilde{\gamma}$ with $X^\top \tilde{\gamma} = 0$ and $\tilde{\gamma}_1 \neq 0$. Then for any $a \in \mathbb{R}$,

$$(T_{\mathbf{X}}(\mathbf{Y}) \mid \mathbf{X}, \gamma) \overset{d}{=} (T_{\mathbf{X}}(\mathbf{Y}) \mid \mathbf{X}, \gamma + a\tilde{\gamma}).$$

By picking $a^* = -\gamma_1/\tilde{\gamma}_1$, we conclude that the power

$$\mathbb{P}(T_{\mathbf{X}}(\mathbf{Y}) = 1 \mid \mathbf{X}, \gamma) = \mathbb{P}(T_{\mathbf{X}}(\mathbf{Y}) = 1 \mid \mathbf{X}, \gamma + a^*\tilde{\gamma}) \leq \alpha$$

by equation (3), since $(\gamma + a^*\tilde{\gamma})_1 = 0$.

On the other hand, we could construct a non-trivial model-X test in the following way. Consider the test statistic

$$T(\mathbf{X}, \mathbf{Y}) = \frac{\mathbf{Y}^\top \mathbf{X}_1}{\|\mathbf{Y}\|} \sim \mathcal{N}(0,1) \text{ if } \beta_1 = 0.$$

where $\mathbf{X}_1$ is the first column of $\mathbf{X}$. If $\beta_1 = 0$, the statistic follows $\mathcal{N}(0,1)$. We will prove the power of the test which rejects when $|T(\mathbf{X}, \mathbf{Y})| > z_{\alpha/2} = \chi_{1,\alpha}$ goes to a constant greater than α for a fixed $\beta_{\text{-}1}$ as $\beta_1 \to \infty$. Let $\varepsilon = \mathbf{Y} - \mathbf{X}\beta$ and note that

$$\frac{\mathbf{X}_1^\top \mathbf{Y}}{\|\mathbf{Y}\|} = \frac{\mathbf{X}_1^\top \mathbf{Y}/\beta_1}{\|\mathbf{Y}\|/\beta_1} = \frac{\|\mathbf{X}_1\|^2 + \sum_{j=2}^p \beta_j \mathbf{X}_1^\top \mathbf{X}_j/\beta_1 + \mathbf{X}_1^\top \varepsilon/\beta_1}{\|\mathbf{X}_1 + \sum_{j=2}^p \beta_j \mathbf{X}_j/\beta_1 + \varepsilon/\beta_1\|} \xrightarrow{p} \|\mathbf{X}_1\| \sim \chi_n.$$

Since $n > 1$, the limit power is greater than α.

C.2 Example where model-X strictly dominates fixed-X

We present a simple example in this section, which reveals that in the finite-sample case, the most powerful test can be model-X instead of fixed-X. We will see in the following sections that this is not the case in asymptotic regimes. Let $X \sim f$ and $Y \mid X \sim \text{Bern}(g_\theta(X))$. Let $(X_i, Y_i)_{i=1}^n$ be i.i.d. copies of (X, Y). Assume $g_0(x) \equiv 1/2$, $g_\theta(x) + g_\theta(-x) \equiv 1$ and $g_\theta(x)$ is an increasing function of x

for $\theta > 0$. We also assume f has symmetric tails; that is, there is a positive constant M such that $X \mid |X| > M \overset{d}{=} -X \mid |X| > M$. Consider testing $H_0 : \theta = 0$ versus $H_1 : \theta = \theta_1 > 0$. Follow the Neyman–Pearson Lemma, the most powerful test is with rejection region of the form

$$\left\{ (X_i, Y_i)_{i=1}^n : \prod_{i=1}^n (\mathbf{1}_{\{Y_i=1\}} g_{\theta_1}(X_i) + \mathbf{1}_{\{Y_i=0\}} g_{\theta_1}(-X_i)) \geq c_\alpha \right\}.$$

For simplicity, let $Z_i = 2Y_i - 1$ be the symmetric version of Y_i, then the rejection region is

$$\left\{ (X_i, Z_i)_{i=1}^n : \prod_{i=1}^n g_{\theta_1}(Z_i X_i) \geq c_\alpha \right\}.$$

Since c_α goes to 1 as $\alpha \to 0$, there is sufficiently small α such that $c_\alpha > g_{\theta_1}(M)$. For this α, it is clear that

$$\{x : \prod_{i=1}^n g_{\theta_1}(z_i x_i) \geq c_\alpha\} \subseteq \{x : |x_i| > M\},$$

for any fixed binary ± 1 sequence $z_1, z_2, \ldots, z_n$. In this region, f is symmetric, so this most powerful test has the correct size α conditional on Z. Put another way, the unique most powerful test is indeed a valid model-X test.

What if we restrict ourselves to fixed-X tests? Due to the discrete nature of this problem, the optimal fixed-X test will involve a randomization step for every level $\alpha \in (0, 1)$ except for a finite number of values. Thus, for almost every α, the most powerful fixed-X test is not the most powerful test.

D Testability of alternative sequences

D.1 Contiguity and q.m.d.

We wish to first note that it is not true that if the alternative sequence is not contiguous to the null then there must exist a test with power converges to one. If the dimension can be fixed, a simple counterexample is Unif$[0, 1]$ versus Unif$[1/2, 3/2]$. If we require them to be the measure on n i.i.d. samples, then let $P_0 = \text{Unif}[0, 1]^n$ and $P_n = \text{Unif}[0, 1 + 1/n]^n$. Obviously, the event $A_n = \{\max_{1 \leq i \leq n} |X_i| > 1\}$ has probability 0 under P_0, but probability $1-(1+1/n)^{-n} \to 1-e^{-1}$ under P_n. So P_n is not contiguous to P_0. The most powerful level-α test is to reject when $\max_{1 \leq i \leq n} |X_i| > 1$ and reject with probability α if $\max |X_i| \leq 1$. The power under P_n is

$$1 - \frac{1}{(1 + 1/n)^n} + \alpha \times \frac{1}{(1 + 1/n)^n} \to 1 - \frac{1 - \alpha}{e} < 1.$$

Now we present some background on contiguity and q.m.d.

Definition 3 (Contiguity, Lehmann and Romano (2006))**.** *Let P_n and Q_n be probability distributions on $(\mathcal{X}_n, \mathcal{F}_n)$. The sequence $\{Q_n\}$ is contiguous to the sequence $\{P_n\}$ if $P_n(E_n) \to 0$ implies $Q_n(E_n) \to 0$ for every sequence $\{E_n\}$ with $E_n \in \mathcal{F}_n$. If $\{Q_n\}$ is contiguous to $\{P_n\}$ and vice versa, we say $\{P_n\}$ and $\{Q_n\}$ are contiguous.*

Lemma 1 (Lehmann and Romano (2006))**.** *Let $\{P_\theta, \theta \in \Omega\}$ with Ω being an open subset of $\mathbb{R}^k$ be* quadratic mean differentiable *(q.m.d.) with densities $p_\theta(\cdot)$. Then for a fixed h, $P^n_{\theta_0 + hn^{-1/2}}$ and $P^n_{\theta_0}$ are contiguous.*

D.2 Total variation distance

Let $H_1 : P \in \mathcal{P}_{1,n}$ be alternatives against $H_0 : P \in \mathcal{P}_{0,n}$, with possibly growing dimensions. The problem is untestable (i.e., every level-α test has power bounded by α) if (Romano 2004)

$$\inf_{P_0 \in \mathcal{P}_{0,n}, P_1 \in \mathcal{P}_{1,n}} \mathrm{TV}(P_0, P_1) = 0.$$

Thus, if

$$\lim_{n\to\infty} \inf_{P_0 \in \mathcal{P}_{0,n}, P_1 \in \mathcal{P}_{1,n}} \mathrm{TV}(P_0, P_1) = 0,$$

the sequence of alternatives is indistinguishable from the null.

To examine the converse, if the total variation distance is lower bounded away from zero, there could still be no test that has non-trivial power against all alternatives. For example, if $\mathcal{P}_{0,n} = \{\mathrm{Unif}[0,1]\}$ and $\mathcal{P}_{1,n} = \{\mathrm{Unif}[0,1/2], \mathrm{Unif}[1/2,1]\}$. For any test ψ which rejects with probability $\psi(x)$ if x is observed,

$$\int_{[0,1]} \psi(x)\, dx \le \alpha.$$

This test cannot have non-trivial power for both alternatives, because at least one inequality holds in

$$\int_{[0,1/2]} \psi(x)\, dx \le \alpha/2, \quad \int_{[1/2,1]} \psi(x)\, dx \le \alpha/2.$$

Another more non-trivial example is testing $n = 0$ against $n \ge 1$ in

$$p_n(x) = 1 + \sin(2n\pi x), x \in [0,1], n \in \mathbb{N}.$$

It is easy to calculate that

$$\mathrm{TV}(p_0, p_n) = 2/\pi, n > 1.$$

But any test level-α test ψ will satisfy

$$\int_{[0,1]} \psi(x)(1 + \sin(2n\pi x))\, dx \le \alpha + \int_{[0,1]} \psi(x) \sin(2n\pi x)\, dx \to \alpha$$

as $n \to \infty$ by Riemann–Lebesgue Lemma.

E Proofs

Lemma 2. *Assume $X \perp\!\!\!\perp Y \mid Z$. Let*

$$\hat{R}_n(t) = \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \le t \mid \mathbf{Y}, \mathbf{Z}).$$

Let $\tilde{\mathbf{X}}$ be a conditionally independent copy of $\mathbf{X}$ given $\mathbf{Y}$ and $\mathbf{Z}$. If

$$(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}), T_n(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z})) \xrightarrow{d} (T, \tilde{T}), \tag{4}$$

where T and $\tilde{T}$ are independent with CDF $R(\cdot)$. Then for every t which is a continuity point of $R(\cdot)$, we have

$$\hat{R}_n(t) \xrightarrow{p} R(t). \tag{5}$$

Proof of Lemma 2. Let t be a continuity point of $R(\cdot)$. By equation (4)

$$\mathbb{E}[\hat{R}_n(t)] = \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t) \to R(t)$$

Now it suffices to show that

$$\mathrm{Var}[\hat{R}_n(t)] \to 0.$$

This is equivalent to

$$\mathbb{E}[\hat{R}_n(t)^2] \to R(t)^2.$$

Note that

$$\begin{aligned}\hat{R}_n(t)^2 &= \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t \mid \mathbf{Y}, \mathbf{Z})^2 \\ &= \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t, T_n(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \leq t \mid \mathbf{Y}, \mathbf{Z}),\end{aligned}$$

hence, also by equation (4),

$$\begin{aligned}\mathbb{E}[\hat{R}_n(t)^2] &= \mathbb{E}[\mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t, T_n(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \leq t \mid \mathbf{Y}, \mathbf{Z})] \\ &= \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t, T_n(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \leq t) \to \mathbb{P}(T \leq t, \tilde{T} \leq t) = R(t)^2.\end{aligned}$$

□

Lemma 3. *Let*

$$\hat{R}_n(t) = \mathbb{P}(T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \leq t \mid \mathbf{Y}, \mathbf{Z}).$$

Suppose for every t which is a continuity point of a CDF $R(\cdot)$, we have

$$\hat{R}_n(t) \xrightarrow{p} R(t). \tag{6}$$

Let $r(1-\alpha) = \inf\{t : R(t) \geq 1-\alpha\}$; suppose $R(\cdot)$ is continuous and strictly increasing at $r(1-\alpha)$, then

$$\hat{r}_n(1-\alpha) \xrightarrow{p} r(1-\alpha).$$

Proof of Lemma 3. This is a direct consequence of Lemma 11.2.1 (ii) in Lehmann and Romano (2006). □

Theorem 9. *The asymptotic unconditional power of the test in Appendix B under the local alternatives is*

$$1 - \Phi\left(z_{1-\alpha} - h\sqrt{\mathrm{Var}_{H_0}(s(X_i, Y_i, Z_i, \theta_0))}\,\mathrm{Corr}_{H_0}\left(\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i), s(X_i, Y_i, Z_i, \theta_0)\right)\right),$$

where,

$$e_0(Y, Z) = \mathbb{E}_{H_0}[\psi(X, Y, Z) \,|\, Y, Z]$$

and s is the score function, which, under very general regularity conditions,[6] *admits the common form*

$$s(X, Y, Z, \theta_0) = \frac{\frac{\partial}{\partial\theta}\big|_{\theta=\theta_0} p_\theta(X, Y, Z)}{p_{\theta_0}(X, Y, Z)},$$

where p_θ is the density of P_θ.

[6] See Theorem 12.2.1 in Lehmann and Romano (2006) for an example of such conditions. There, the notation $\tilde{\eta}$ is used instead of s.

Proof of Theorem 9. Consider an asymptotically linear statistic

$$T_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = n^{-1/2} \sum_{i=1}^{n} \psi(X_i, Y_i, Z_i) + o_{\mathbb{P}_{H_0}}(1),$$

Suppose we know the direction of the alternative and thus would like a test that rejects when T_n is above a threshold. Since the test is to be valid conditional on $(\mathbf{Y}, \mathbf{Z})$, it would be equivalent to consider the statistic

$$S_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = T_n - n^{-1/2} \sum_{i=1}^{n} e_0(Y_i, Z_i) = n^{-1/2} \sum_{i=1}^{n} (\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i)) + o_{\mathbb{P}_{H_0}}(1),$$

where

$$e_0(y, z) = \mathbb{E}_{H_0}[\psi(X_i, Y_i, Z_i) \mid Y_i = y, Z_i = z].$$

Under the null,

$$S_n \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}_{H_0}[v_0(Y, Z)]\right),$$

where

$$v_0(y, z) = \mathrm{Var}_{H_0}[\psi(X_i, Y_i, Z_i) \mid Y_i = y, Z_i = z].$$

In addition, note that if $\tilde{\mathbf{X}}$ is a copy of $\mathbf{X}$ conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ (as in Lemma 2), then

$$\begin{aligned}
&\mathrm{Cov}_{H_0}(\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i), \psi(\tilde{X}_i, Y_i, Z_i) - e_0(Y_i, Z_i)) \\
&\quad = \mathbb{E}_{H_0}[\mathrm{Cov}_{H_0}(\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i), \psi(\tilde{X}_i, Y_i, Z_i) - e_0(Y_i, Z_i) \mid Y_i, Z_i)] \\
&\qquad + \mathrm{Cov}_{H_0}(\mathbb{E}_{H_0}[\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i) \mid Y_i, Z_i], \mathbb{E}_{H_0}[\psi(\tilde{X}_i, Y_i, Z_i) - e_0(Y_i, Z_i) \mid Y_i, Z_i]) \\
&\quad = 0 + 0 = 0.
\end{aligned}$$

By the bivariate central limit theorem, under H_0,

$$\begin{pmatrix} S_n(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \\ S_n(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \mathbb{E}_{H_0}[v_0(Y, Z)] & 0 \\ 0 & \mathbb{E}_{H_0}[v_0(Y, Z)] \end{bmatrix} \right).$$

The test ϕ_n rejects when $S_n > \hat{r}_n(1-\alpha)$, accepts when $S_n < \hat{r}_n(1-\alpha)$, and possibly randomizes when $S_n = \hat{r}_n(1-\alpha)$. By Lemma 3, $\hat{r}_n(1-\alpha) \xrightarrow{p} z_{1-\alpha}\sqrt{\mathbb{E}_{H_0}[v_0(Y, Z)]}$ under H_0.

Since the null distribution $\mathbb{P}_{H_0} = P^n_{\theta_0}$ and the alternative is a sequence $P^n_{\theta_0 + hn^{-1/2}}$, where the family is q.m.d., by contiguity (Lemma 1), $\hat{r}_n(1-\alpha) \xrightarrow{p} z_{1-\alpha}\sqrt{\mathbb{E}_{H_0}[v_0(Y, Z)]}$ under the alternative sequence as well. To study the asymptotic power under local alternatives, we introduce Le Cam's Third Lemma.

Lemma 4 (Le Cam's Third Lemma, Corollary 12.3.2 in Lehmann and Romano (2006))**.** *If*

$$\begin{pmatrix} X_n \\ \log \frac{dQ_n}{dP_n} \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} \\ \sigma_{1,2} & \sigma_2^2 \end{bmatrix} \right) \text{ under } P_n,$$

where $\frac{dQ_n}{dP_n}$ *is the likelihood ratio and* $\mu_2 = -\sigma_2^2/2$ *so that* Q_n *is contiguous to* P_n*, then*

$$X_n \xrightarrow{d} \mathcal{N}(\mu_1 + \sigma_{1,2}, \sigma_1^2) \text{ under } Q_n.$$

By taking X_n to be S_n, P_n to be $P^n_{\theta_0}$ and Q_n to be $P^n_{\theta_0+hn^{-1/2}}$ in Lemma 4, under $P^n_{\theta_0+hn^{-1/2}}$, $S_n \xrightarrow{d} \mathcal{N}(\sigma_{1,2}, \mathbb{E}_{H_0}[v_0(Y,Z)])$ ($\log \frac{dQ_n}{dP_n}$ is asymptotically the score; see Example 12.3.8 in Lehmann and Romano (2006)), where

$$\sigma_{1,2} = h \operatorname{Cov}_{H_0}(\psi(X_i, Y_i, Z_i) - e_0(Y_i, Z_i), s(X_i, Y_i, Z_i, \theta_0)).$$

The asymptotic power is thus

$$1 - \Phi\left(z_{1-\alpha} - \frac{\sigma_{1,2}}{\sqrt{\mathbb{E}_{H_0}[v_0(Y,Z)]}}\right).$$

□

Theorem 1. *In Setting 1, the CRT with T_{MC} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\sqrt{\sigma^2 + v_Z^2}}.$$

Proof of Theorem 1. We only prove the one-sided case, while the two-sided case can be dealt with almost identically.

Under the null, $T_{\mathrm{MC}}(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{Z} \sim \mathcal{N}(n^{-1}\mathbf{Y}^\top \mathbf{Z}\xi, \|\mathbf{Y}\|^2/n^2)$, so the power is

$$\mathbb{P}_{\beta=h/\sqrt{n}}\left(\frac{1}{n}\mathbf{Y}^\top \mathbf{X} \geq \frac{1}{n}\mathbf{Y}^\top \mathbf{Z}\xi + z_{1-\alpha}\frac{\|\mathbf{Y}\|}{n}\right) = \mathbb{P}_{\beta=h/\sqrt{n}}\left(\frac{1}{\sqrt{n}}\mathbf{Y}^\top(\mathbf{X} - \mathbf{Z}\xi) \geq z_{1-\alpha}\frac{\|\mathbf{Y}\|}{\sqrt{n}}\right). \quad (7)$$

The elements of $\mathbf{X} - \mathbf{Z}\xi$ are conditionally independent given $\mathbf{Y}$ and the distribution $\mathcal{L}(X - Z^\top\xi \,|\, Y)$ is (by applying the conditional distribution formula to the bivariate Gaussian distribution of $(X - Z^\top\xi, Y)$)

$$\mathcal{N}\left(\frac{\beta Y}{(\theta+\beta\xi)^\top\Sigma(\theta+\beta\xi) + \beta^2 + \sigma^2}, 1 - \frac{\beta^2}{(\theta+\beta\xi)^\top\Sigma(\theta+\beta\xi) + \beta^2 + \sigma^2}\right).$$

Thus,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\mathbf{Y}^\top(\mathbf{X} - \mathbf{Z}\xi) \mid \mathbf{Y} \\
&\qquad \sim \mathcal{N}\left(\frac{n^{-1/2}\beta\|\mathbf{Y}\|^2}{(\theta+\beta\xi)^\top\Sigma(\theta+\beta\xi) + \beta^2 + \sigma^2}, \frac{\|\mathbf{Y}\|^2}{n}\left(1 - \frac{\beta^2}{(\theta+\beta\xi)^\top\Sigma(\theta+\beta\xi) + \beta^2 + \sigma^2}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}_{\beta=h/\sqrt{n}}\left(\frac{1}{\sqrt{n}}\mathbf{Y}^\top(\mathbf{X} - \mathbf{Z}\xi) \geq z_{1-\alpha}\frac{\|\mathbf{Y}\|}{\sqrt{n}}\right) \\
&= \mathbb{E}\left[\mathbb{P}_{\beta=h/\sqrt{n}}\left(\frac{1}{\sqrt{n}}\mathbf{Y}^\top(\mathbf{X} - \mathbf{Z}\xi) \geq z_{1-\alpha}\frac{\|\mathbf{Y}\|}{\sqrt{n}} \,|\, \mathbf{Y}\right)\right] \\
&= \mathbb{E}\left[\Phi\left(\frac{\frac{n^{-1}h\|\mathbf{Y}\|^2}{(\theta+h\xi/\sqrt{n})^\top\Sigma(\theta+h\xi/\sqrt{n}) + h^2/n + \sigma^2} - z_{1-\alpha}\frac{\|\mathbf{Y}\|}{\sqrt{n}}}{\frac{\|\mathbf{Y}\|}{\sqrt{n}}\sqrt{1 - \frac{h^2/n}{(\theta+h\xi/\sqrt{n})^\top\Sigma(\theta+h\xi/\sqrt{n}) + h^2/n + \sigma^2}}}\right)\right] \\
&\to \Phi\left(\frac{h}{\sqrt{v_Z^2 + \sigma^2}} - z_{1-\alpha}\right),
\end{aligned}$$

where we used $\|\mathbf{Y}\|^2/n \xrightarrow{p} v_Z^2 + \sigma^2$. To see why this is the case, note that

$$Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{h^2}{n} + (\theta + h\xi/\sqrt{n})^\top \Sigma_Z (\theta + h\xi/\sqrt{n}) + \sigma^2\right),$$

so we only need to show

$$(\theta + h\xi/\sqrt{n})^\top \Sigma_Z (\theta + h\xi/\sqrt{n}) \to v_Z^2. \tag{8}$$

Equation (8) holds because by assumption, $\theta^\top \Sigma_Z \theta \to v_Z^2$, $n^{-1}\xi^\top \Sigma_Z \xi \to 0$, and by the Cauchy–Schwarz inequality, the cross term satisfies

$$\theta^\top \Sigma_Z \xi/\sqrt{n} \le \sqrt{n^{-1}\theta^\top \Sigma_Z \theta \cdot \xi^\top \Sigma_Z \xi} \to 0.$$

□

Theorem 2. *In Setting 1 with $\kappa < 1$, the CRT with T_{OLS} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\sigma}\sqrt{1-\kappa}.$$

Proof of Theorem 2. We only prove the one-sided case, while the two-sided case can be dealt with almost identically.

We look at the expression of the normalized OLS statistic $T_{\text{OLS}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \sqrt{n}\hat{\beta}$:

$$T_{\text{OLS}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \sqrt{n}\hat{\beta} = \frac{\mathbf{X}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/\sqrt{n}}{\mathbf{X}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{X}/n}, \tag{9}$$

and the rejection region is $\{T_{\text{OLS}} \ge \hat{c}_\alpha\}$, where $\hat{c}_\alpha$ is the upper α-quantile of the distribution of

$$\tilde{T}_{\text{OLS}}(\tilde{\mathbf{X}}, \mathbf{Y}, \mathbf{Z}) = \frac{\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/\sqrt{n}}{\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\tilde{\mathbf{X}}/n},$$

conditional on $(\mathbf{Y}, \mathbf{Z})$. Looking at the numerator and denominator individually, we see that

$$\mathcal{L}\left(\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/\sqrt{n} \mid \mathbf{Y}, \mathbf{Z}\right) \sim \mathcal{N}(0, \mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/n),$$
$$\mathcal{L}\left(\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\tilde{\mathbf{X}}/n \mid \mathbf{Y}, \mathbf{Z}\right) \sim n^{-1}\chi^2_{n-p} \text{ (Cochran 1934, Cochran's Theorem).}$$

Now we assume we are under the local alternative $\beta = h/\sqrt{n}$. Again by Cochran's Theorem,

$$\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/n \sim n^{-1}(\sigma^2 + \beta^2)\chi^2_{n-p}.$$

Thus, for any $t \in \mathbb{R}$,

$$\mathbb{P}\left(\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y}/\sqrt{n} \le t \mid \mathbf{Y}, \mathbf{Z}\right) \xrightarrow{p} \Phi(t/\sqrt{\sigma^2(1-\kappa)}).$$

On the other hand, $\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\tilde{\mathbf{X}}/n \perp\!\!\!\perp (\mathbf{Y}, \mathbf{Z})$ and for any $t \ne 1 - \kappa$,

$$\mathbb{P}\left(\tilde{\mathbf{X}}^\top (I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1}\mathbf{Z}^\top)\tilde{\mathbf{X}}/n \le t \mid \mathbf{Y}, \mathbf{Z}\right) \to \mathbf{1}_{\{t > 1-\kappa\}}.$$

By Lemma 5, for any $t \in \mathbb{R}$,

$$\mathbb{P}\left(\tilde{T}_{\mathrm{OLS}} \le t \mid \mathbf{Y}, \mathbf{Z}\right) \xrightarrow{p} \Phi(\sqrt{1-\kappa}t/\sigma),$$

and by Lemma 11.2.1 (ii) in Lehmann and Romano (2006),

$$\hat{c}_\alpha \xrightarrow{p} z_{1-\alpha}\frac{\sigma}{\sqrt{1-\kappa}}.$$

On the other hand, we have the test statistic itself satisfies

$$T_{\mathrm{OLS}} \mid \mathbf{X}, \mathbf{Z} \sim \mathcal{N}(h, \sigma^2 n\hat{\Omega}_{11}),$$

where $\hat{\Omega}$ is the inverse of the matrix $(\mathbf{X}, \mathbf{Z})^\top(\mathbf{X}, \mathbf{Z})$ that follows an inverse-Wishart distribution, and then $n\hat{\Omega}_{11} \xrightarrow{p} 1/(1-\kappa)$ by moment calculations. Therefore, $T_{\mathrm{OLS}} \xrightarrow{d} \mathcal{N}(h, \sigma^2/(1-\kappa))$. It follows then

$$\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\mathrm{OLS}} \ge \hat{c}_\alpha\right) \to 1 - \Phi_{\sigma^2/(1-\kappa)}(\frac{\sigma}{\sqrt{1-\kappa}}z_{1-\alpha} - h) = \Phi\left(\frac{h}{\sigma}\sqrt{1-\kappa} - z_{1-\alpha}\right),$$

where $\Phi_{\sigma^2/(1-\kappa)}$ is the CDF of $\mathcal{N}(0, \sigma^2/(1-\kappa))$.

□

Lemma 5. *Let $\mathcal{L}(X_n \,|\, Z_n)$ have random CDF F_n and $\mathcal{L}(Y_n \,|\, Z_n)$ have deterministic CDF G_n (in other words, $Y_n \perp\!\!\!\perp Z_n$). Let $\mathcal{L}(Y_n \,|\, Z_n)$ converge in distribution to a point mass at c, $c > 0$, and for a continuous and deterministic CDF F on $\mathbb{R}$, let $F_n(t) \xrightarrow{p} F(t)$ for any $t \in \mathbb{R}$. Let H_n be the CDF of $\mathcal{L}(X_nY_n \,|\, Z_n)$. Then for any $t \in \mathbb{R}$, $H_n(t) \xrightarrow{p} F(t/c)$.*

Proof of Lemma 5. Without loss of generality, assume $c = 1$. Fix $t \in \mathbb{R}$ and $\varepsilon > 0$. Pick $\delta > 0$ such that $|F(t) - F(t/(1+\delta))| \le \varepsilon/2$.

$$\begin{aligned}
H_n(t) &= \mathbb{P}\left(X_nY_n \le t \mid Z_n\right) \\
&\ge \mathbb{P}\left(X_n \le \frac{t}{1+\delta}, Y_n \le 1+\delta \mid Z_n\right) \\
&= \mathbb{P}\left(\left\{X_n \le \frac{t}{1+\delta}\right\} \setminus \{Y_n > 1+\delta\} \mid Z_n\right) \\
&\ge \mathbb{P}\left(X_n \le \frac{t}{1+\delta} \mid Z_n\right) - \mathbb{P}\left(Y_n > 1+\delta \mid Z_n\right) \\
&= F_n(t/(1+\delta)) - (1 - G_n(1+\delta)) \xrightarrow{p} F(t/(1+\delta)).
\end{aligned}$$

It follows that $\mathbb{P}(H_n(t) \ge F(t/(1+\delta)) - \varepsilon/2) \to 1$. By the choice of δ, $\mathbb{P}(H_n(t) \ge F(t) - \varepsilon) \to 1$. Similarly, we can get $\mathbb{P}(H_n(t) \le F(t) + \varepsilon) \to 1$, thus proving the claim. □

Theorem 3. *Under Setting 1 with $\Sigma_Z = I$ and $\xi = 0$, if the empirical distribution of $(\sqrt{n}\theta_j)_{j=1}^{p-1}$ converges to a distribution represented by a random variable B_0 and $\|\sqrt{n}\theta\|_2^2/p \to \mathbb{E}[B_0^2]$, then the CRT with the distilled lasso statistic with lasso parameter λ has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h}{\tau_\lambda}.$$

Proof of Theorem 3. To use the results in Bayati and Montanari (2011), we apply the following re-normalization: assume $(\mathbf{X}, \mathbf{Z})$ is divided by $\sqrt{n}$, (β, θ) is multiplied by $\sqrt{n}$, and the statistic is $T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda)^\top \mathbf{X}$. The proof is a direct consequence of Lemma 6. □

Lemma 6. *Assume Setting 1 with $\Sigma_Z = I$, $\xi = 0$, $\sqrt{n}\beta$ universally bounded (but not necessarily a constant), and ε_i's and X_i's do not change with n, p as long as $n \geq i$. If the empirical distribution of $(\sqrt{n}\theta_j)_{j=1}^{p-1}$ converges to a distribution represented by a random variable B_0 and $\|\sqrt{n}\theta\|_2^2/p \to \mathbb{E}[B_0^2]$, then we have*

$$\frac{\|\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda\|_2^2}{n} \overset{\text{a.s.}}{\to} \frac{\lambda^2}{\alpha_\lambda^2}, \tag{10}$$

$$\frac{1}{n}(\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda)^\top (\mathbf{Y} - \mathbf{Z}\theta) \overset{\text{a.s.}}{\to} \frac{\lambda}{\alpha_\lambda \tau_\lambda}\sigma^2, \tag{11}$$

and

$$T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) - \frac{\sqrt{n}\beta\lambda}{\alpha_\lambda \tau_\lambda} \overset{d}{\to} \mathcal{N}\left(0, \frac{\lambda^2}{\alpha_\lambda^2}\right).$$

Here, α_λ and τ_λ satisfy

$$\begin{aligned} \lambda &= \alpha_\lambda \tau_\lambda \left(1 - \kappa \mathbb{E}[\eta'(B_0 + \tau_\lambda W; \alpha_\lambda \tau_\lambda)]\right), \\ \tau_\lambda^2 &= \sigma^2 + \kappa \mathbb{E}[(\eta(B_0 + \tau_\lambda W; \alpha_\lambda \tau_\lambda) - B_0)^2], \end{aligned} \tag{12}$$

where $W \sim \mathcal{N}(0, 1)$ is independent of B_0 and η' is the derivative of η.

Proof of Lemma 6. We assume ε_i's and X_i's do not change with n, p to satisfy Definition 1 (b) in Bayati and Montanari (2011) by

$$\|\varepsilon\|_2^2/n \overset{\text{a.s.}}{\to} \sigma^2 \text{ and } \|\beta\mathbf{X}\|_2^2/n \overset{\text{a.s.}}{\to} 0.$$

This additional assumption on ε_i's and X_i's does not change the asymptotic power; in fact, it does not change the power for any fixed pair of (n, p), because the power is a marginal quantity for each pair of (n, p) and does not depend on the relationship of the random variables between different pairs of (n, p)'s.

To use the results in Bayati and Montanari (2011), we again apply the following re-normalization: assume $(\mathbf{X}, \mathbf{Z})$ is divided by $\sqrt{n}$, (β, θ) is multiplied by $\sqrt{n}$, and the statistic is $T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda)^\top \mathbf{X}$.

Note that in the $\mathbf{Y}$ against $\mathbf{Z}$ regression, we can absorb $\mathbf{X}$ into the error and under the assumption that β stays universally bounded, the effective error

$$\varepsilon' = \mathbf{Y} - \mathbf{Z}\theta = \varepsilon + \beta\mathbf{X}$$

still has the property that its empirical distribution converges to $\mathcal{N}(0, \sigma^2)$ and its second moment converges to σ^2. We first prove (10). The AMP iteration is

$$\begin{aligned} \theta^{t+1} &= \eta(\mathbf{Z}^\top z^t + \theta^t; \alpha_\lambda \tau_t), \\ z^t &= \mathbf{Y} - \mathbf{Z}\theta^t + \kappa z^{t-1} \langle \eta'(\mathbf{Z}^\top z^{t-1} + \theta^{t-1}; \alpha_\lambda \tau_{t-1}) \rangle \\ \tau_{t+1}^2 &= \sigma^2 + \kappa \mathbb{E}[(\eta(B_0 + \tau_t W; \alpha_\lambda \tau_t) - B_0)^2], \end{aligned}$$

where η' is the derivative of η and $\langle \cdot \rangle$ means taking the average of the coordinates of a vector. We denote

$$w_t = \kappa \langle \eta'(\mathbf{Z}^\top z^{t-1} + \theta^{t-1}; \alpha_\lambda \tau_{t-1}) \rangle.$$

We first see that by the reverse triangle inequality,

$$\left| \frac{\|\mathbf{Y} - \mathbf{Z}\hat{\theta}\|_2}{\sqrt{n}} - \frac{\|\mathbf{Y} - \mathbf{Z}\theta^t\|_2}{\sqrt{n}} \right| \leq \frac{\|\mathbf{Z}(\theta^t - \hat{\theta})\|_2}{\sqrt{n}}.$$

Note that

$$\frac{\|\mathbf{Z}(\theta^t - \hat{\theta})\|_2^2}{n} \leq \frac{\sigma_{\max}^2(\mathbf{Z})\|\theta^t - \hat{\theta}\|_2^2}{n},$$

where $\sigma_{\max}(\mathbf{Z})$ is almost surely bounded (see, e.g., Theorem F.2 in Bayati and Montanari (2011)) and Theorem 1.8 in Bayati and Montanari (2011) states that

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{\|\theta^t - \hat{\theta}\|_2^2}{n} = 0 \text{ almost surely.}$$

Thus,

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{\|\mathbf{Z}(\theta^t - \hat{\theta})\|_2^2}{n} = 0 \text{ almost surely.}$$

Now we just have to show

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{\|\mathbf{Y} - \mathbf{Z}\theta^t\|_2^2}{n} = \frac{\lambda^2}{\alpha_\lambda^2} \text{ almost surely,} \tag{13}$$

which will prove (10). By definition, $\mathbf{Y} - \mathbf{Z}\theta^t = z^t - z^{t-1}w_t$. By the reverse triangle inequality,

$$\left| \frac{\|z^t - z^{t-1}w_t\|_2}{\sqrt{n}} - \frac{\|z^{t-1}(1-w_t)\|_2}{\sqrt{n}} \right| \leq \frac{\|z^t - z^{t-1}\|_2}{\sqrt{n}},$$

and the right hand side goes to 0 as stated by Lemma 4.3 in Bayati and Montanari (2011). Thus, to prove (13), we can just analyze the limit of $\|z^{t-1}(1-w_t)\|_2^2/n$. Directly by Lemma 4.1 in Bayati and Montanari (2011),

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{1}{n}\|z^{t-1}\|_2^2 = \lim_{t\to\infty} \tau_t^2 = \tau_\lambda^2 \text{ almost surely.}$$

Almost surely,

$$\begin{aligned}
\lim_{t\to\infty}\lim_{n\to\infty} w_t &= \lim_{t\to\infty} \kappa\mathbb{E}[\eta'(B_0 + \tau_{t-1}W; \alpha_\lambda\tau_{t-1})] && \text{(Equation (4.11) in Bayati and Montanari (2011))} \\
&= \kappa\mathbb{E}[\eta'(B_0 + \tau_\lambda W; \alpha_\lambda\tau_\lambda)] && \text{(bounded convergence theorem)} \\
&= 1 - \frac{\lambda}{\alpha_\lambda\tau_\lambda}. && \text{(definition of } \alpha_\lambda \text{ and } \tau_\lambda\text{, Equation (12))}
\end{aligned} \tag{14}$$

Combining the above results, we get

$$\lim_{t\to\infty}\lim_{n\to\infty} \frac{\|z^{t-1}(1-w_t)\|_2^2}{n} = \lim_{t\to\infty}\lim_{n\to\infty} (1-w_t)^2 \lim_{t\to\infty}\lim_{n\to\infty} \frac{1}{n}\|z^{t-1}\|_2^2 = \frac{\lambda^2}{\alpha_\lambda^2} \text{ almost surely.}$$

Now we prove (11). By (F.12) in Lemma F.3(d) in Bayati and Montanari (2011) (take $\varphi(u,v) = v$, take their r and s to both be t, their w is our ε' and their b^t is our $\varepsilon' - z^t$),

$$\lim_{n\to\infty} \langle \varepsilon' - z^t, \varepsilon' \rangle = 0 \Rightarrow \sigma^2 = \langle \varepsilon', \varepsilon' \rangle = \lim_{n\to\infty} \langle \varepsilon', z^t \rangle \text{ almost surely.}$$

Thus,

$$\langle \varepsilon', \mathbf{Y} - \mathbf{Z}\theta^t \rangle = \langle \varepsilon', z^t - z^{t-1}w^t \rangle \overset{\text{a.s.}}{\to} \sigma^2 - \sigma^2 w^t \text{ as } n \to \infty.$$

Combining the above equation with (14), we see that

$$\lim_{t\to\infty} \lim_{n\to\infty} \langle \varepsilon', \mathbf{Y} - \mathbf{Z}\theta^t \rangle = \sigma^2 \frac{\lambda}{\alpha_\lambda \tau_\lambda} \text{ almost surely.}$$

What we are interested in is the limit of $\langle \varepsilon', \mathbf{Y} - \mathbf{Z}\hat{\theta} \rangle$ as $n \to \infty$. Note that

$$\langle \varepsilon', \mathbf{Y} - \mathbf{Z}\theta^t \rangle - \langle \varepsilon', \mathbf{Y} - \mathbf{Z}\hat{\theta} \rangle = \langle \varepsilon', \mathbf{Z}(\hat{\theta} - \theta^t) \rangle.$$

By the Cauchy–Schwartz inequality,

$$|\langle \varepsilon', \mathbf{Z}(\hat{\theta} - \theta^t) \rangle| \le \sqrt{\frac{\|\varepsilon'\|_2^2}{n} \frac{\|\mathbf{Z}(\theta^t - \hat{\theta})\|_2^2}{n}}.$$

Since $\|\varepsilon'\|_2^2/n \overset{\text{a.s.}}{\to} \sigma^2$ and we have showed $\|\mathbf{Z}(\theta^t - \hat{\theta})\|_2^2/n \overset{\text{a.s.}}{\to} 0$ (as $n \to \infty$ then $t \to \infty$), this means

$$\lim_{n\to\infty} \langle \varepsilon', \mathbf{Y} - \mathbf{Z}\hat{\theta} \rangle = \lim_{t\to\infty} \lim_{n\to\infty} \langle \varepsilon', \mathbf{Y} - \mathbf{Z}\theta^t \rangle = \sigma^2 \frac{\lambda}{\alpha_\lambda \tau_\lambda} \text{ almost surely.}$$

Note that

$$\mathbf{X} \mid \mathbf{Y}, \mathbf{Z} \sim \mathcal{N}\left(\frac{\beta}{n\sigma^2 + \beta^2} \varepsilon', \frac{\sigma^2}{n\sigma^2 + \beta^2} \right),$$

where we remind the reader that $\varepsilon' = \mathbf{Y} - \mathbf{Z}^\top \theta = \varepsilon + \beta \mathbf{X}$. Hence,

$$T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}, \mathbf{Z} \sim \mathcal{N}\left(\frac{\beta}{n\sigma^2 + \beta^2} (\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda)^\top \varepsilon', \frac{\sigma^2}{n\sigma^2 + \beta^2} \|\mathbf{Y} - \mathbf{Z}\hat{\theta}_\lambda\|_2^2 \right)$$

Now it is clear that

$$T_{\text{distilled}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) - \frac{\beta\lambda}{\alpha_\lambda \tau_\lambda} \overset{d}{\to} \mathcal{N}\left(0, \frac{\lambda^2}{\alpha_\lambda^2}\right).$$

□

Theorem 4. *In Setting 1 with ξ and $\mathrm{Var}(X \mid Z)$ unknown but fixed to be 1, if there are m additional data points $(X_i, Z_i)_{i=n+1}^{n+m}$, $n_* = n + m$, $n/n_* \to \kappa_*$ and $\kappa\kappa_* < 1$, then the conditional CRT with statistic T_{MC} has asymptotic power lower-bounded (the* $\liminf$ *is lower-bounded) by that of a z-test with standardized effect size*

$$\frac{h\sqrt{1 - \kappa\kappa_*}}{\sqrt{\sigma^2 + v_Z^2 \frac{1}{1 - \kappa\kappa_*}}}$$

and upper-bounded (the $\limsup$ *is upper-bounded) by that of a z-test with standardized effect size*

$$\frac{h\sqrt{1 - \kappa\kappa_*}}{\sqrt{\sigma^2 + v_Z^2 \max\left(0, \frac{1 - \frac{(1+\sqrt{1/\kappa})^2}{(1-\sqrt{\kappa\kappa_*})^2}\kappa\kappa_*}{1 - \kappa\kappa_*}\right)}}.$$

Proof of Theorem 4. This proof uses some results and notation in the detailed introduction of the conditional CRT in Appendix G and should be read after that.

We first show that in our asymptotic regime, we can assume $\mathrm{Var}(X \mid Z)$ is known. Then we analyze the asymptotic power assuming $\mathrm{Var}(X \mid Z)$ is known.

Knowledge of $\mathrm{Var}(X \mid Z)$**.** We show that we could assume $\mathrm{Var}(X \mid Z)$ is known by making the following claim: suppose we obtain a cutoff without knowing $\mathrm{Var}(X \mid Z)$ and another oracle cutoff with the knowledge of $\mathrm{Var}(X \mid Z)$ and then we proceed to use the two cutoffs to perform the CRT with the same test statistic. The two decisions differ if and only if the test statistic falls between the two cutoffs, and we claim the probability of this happening goes to 0.

By the conditional nature, the following modified statistic is equivalent when used for the CRT.

$$T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) = \frac{T^{\text{ess}}_{\text{MC}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*)}{\sqrt{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top I_{n_*\times n}\mathbf{Y}}} = \frac{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top \varepsilon_*^X}{\sqrt{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top I_{n_*\times n}\mathbf{Y}}}.$$

If we know $\mathrm{Var}(X \mid Z) = 1$, the test is simply $T_{\text{modified}} \geq z_{1-\alpha}$. When we do not know $\mathrm{Var}(X \mid Z)$, the test can be done by replacing $z_{1-\alpha}$ with the α-upper quantile of

$$\mathcal{L}\left(\frac{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\| \frac{W}{\|W\|}}{\sqrt{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top I_{n_*\times n}\mathbf{Y}}} \mid \mathbf{Y}, \mathbf{Z}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|\right), W \text{ is independent } \mathcal{N}\left(0, I_{n_*-p}\right),$$

which we denote by $\hat{c}_\alpha^n$. Evidently, we are interested in the limiting behavior of

$$\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (\min(z_{1-\alpha}, \hat{c}_\alpha^n), \max(z_{1-\alpha}, \hat{c}_\alpha^n))\right),$$

which we will show goes to 0.

Since

$$\mathcal{L}\left(\frac{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} \frac{\|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|}{\sqrt{n_*-p}} W}{\sqrt{\mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top I_{n_*\times n}\mathbf{Y}}} \mid \mathbf{Y}, \mathbf{Z}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|\right) = \mathcal{N}\left(0, \underbrace{\frac{\|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|^2}{n_*-p}}_{\xrightarrow{p} 1}\right)$$

and

$$\mathcal{L}\left(\frac{\sqrt{n_*-p}}{\|W\|} \mid \mathbf{Y}, \mathbf{Z}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|\right) = \underbrace{\sqrt{n_*-p} \cdot \text{Inv-}\chi^2_{n_*-p}}_{\xrightarrow{p} 1},$$

we can use Lemma 5 and Lemma 11.2.1 (ii) in Lehmann and Romano (2006) to establish that $\hat{c}_\alpha^n$ converges to $z_{1-\alpha}$ in probability (note that by this analysis, the statement is true under both the null and the alternative sequence). Now for any $\delta > 0$,

$$\begin{aligned}\{T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (\min(z_{1-\alpha}, \hat{c}_\alpha^n), \max(z_{1-\alpha}, \hat{c}_\alpha^n))\} \\ \subseteq \{|z_{1-\alpha} - \hat{c}_\alpha^n| > \delta\} \cup \{T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (z_{1-\alpha} - \delta, z_{1-\alpha} + \delta)\}.\end{aligned}$$

Under $\beta = h/\sqrt{n}$, by calculating $\mathcal{L}(\varepsilon_*^X \mid \mathbf{Y}, \mathbf{Z}_*)$, we get

$$\begin{aligned}T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \mid \mathbf{Y}, \mathbf{Z}_* &\sim \mathcal{N}(\mu_n(\mathbf{Y}, \mathbf{Z}_*), \sigma_n^2(\mathbf{Y}, \mathbf{Z}_*)), \text{ where} \\ \mu_n(\mathbf{Y}, \mathbf{Z}_*) &= \frac{h}{\sigma^2 + h^2/n} \frac{\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta))}{\sqrt{n}\sqrt{\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Y}}}, \\ \sigma_n^2(\mathbf{Y}, \mathbf{Z}_*) &= \frac{\sigma^2}{\sigma^2 + h^2/n} + \frac{h^2/n}{h^2/n + \sigma^2} \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Y}} \\ &\quad - \frac{h^2/n}{\sigma^2 + h^2/n} \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Y}}.\end{aligned}$$

Note that

$$\begin{aligned}
&\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} \\
&= \mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top \mathbf{Z}_*(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} - \mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} \\
&= \mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}(\mathbf{Z}_*^\top \mathbf{Z}_* - \mathbf{Z}^\top \mathbf{Z})(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} \\
&= \mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\left(\sum_{i=n+1}^{n+m} \mathbf{Z}_i \mathbf{Z}_i^\top\right)(\mathbf{Z}_*^\top \mathbf{Z})^{-1}\mathbf{Z}^\top \mathbf{Y} \geq 0,
\end{aligned}$$

where $\mathbf{Z}_i$ is the ith row of $\mathbf{Z}_*$ as a column vector. Therefore, we see that $\sigma_n^2(\mathbf{Y}, \mathbf{Z}_*) \geq \sigma^2/(\sigma^2 + h^2/n)$, so the conditional density of $\mathcal{L}(T_{\text{modified}} \mid \mathbf{Y}, \mathbf{Z}_*)$ is upper bounded by $\sqrt{\sigma^2 + h^2/n}/\sqrt{2\pi\sigma^2}$. Thus,

$$\begin{aligned}
&\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (z_{1-\alpha} - \delta, z_{1-\alpha} + \delta)\right) \\
&= \mathbb{E}_{\beta=h/\sqrt{n}}\left[\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (z_{1-\alpha} - \delta, z_{1-\alpha} + \delta) \mid \mathbf{Y}, \mathbf{Z}_*\right)\right] \\
&\leq \mathbb{E}_{\beta=h/\sqrt{n}}\left[\frac{2\delta\sqrt{\sigma^2 + h^2/n}}{\sqrt{2\pi\sigma^2}}\right] \leq \frac{2\delta\sqrt{\sigma^2 + h^2/n}}{\sqrt{2\pi\sigma^2}}.
\end{aligned}$$

We then obtain

$$\begin{aligned}
&\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (\min(z_{1-\alpha}, \hat{c}_\alpha^n), \max(z_{1-\alpha}, \hat{c}_\alpha^n))\right) \\
&\quad \leq \underbrace{\mathbb{P}_{\beta=h/\sqrt{n}}\left(|z_{1-\alpha} - \hat{c}_\alpha^n| > \delta\right)}_{\to 0} + \underbrace{\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (z_{1-\alpha} - \delta, z_{1-\alpha} + \delta)\right)}_{\leq \frac{2\delta\sqrt{\sigma^2+h^2/n}}{\sqrt{2\pi\sigma^2}} \to \frac{2\delta}{\sqrt{2\pi}}}
\end{aligned}$$

and hence

$$\limsup \mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{modified}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \in (\min(z_{1-\alpha}, \hat{c}_\alpha^n), \max(z_{1-\alpha}, \hat{c}_\alpha^n))\right) \leq \frac{2\delta}{\sqrt{2\pi}}.$$

Let $\delta \to 0$ and the claim is proved.

Analysis of power assuming $\mathrm{Var}(X \mid Z) = 1$ is known. Since we condition on $\mathbf{Y}$ in the model-X framework, it would be equivalent to consider

$$T_{\text{model-X}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) = \mathbf{Y}^\top(\mathbf{X} - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top \mathbf{X}_*)/\|\mathbf{Y}\|.$$

By straightforward calculation,

$$\begin{aligned}
T_{\text{model-X}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \mid \mathbf{Y}, \mathbf{Z}_* &\sim \mathcal{N}(\mu_\beta(\mathbf{Y}, \mathbf{Z}_*), \sigma_\beta^2(\mathbf{Y}, \mathbf{Z}_*)), \text{ where} \\
\mu_\beta(\mathbf{Y}, \mathbf{Z}_*) &= \frac{\beta}{\sigma^2 + \beta^2}\frac{\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta))}{\|\mathbf{Y}\|}, \\
\sigma_\beta^2(\mathbf{Y}, \mathbf{Z}_*) &= \frac{\sigma^2}{\sigma^2 + \beta^2} + \frac{\beta^2 - \sigma^2}{\beta^2 + \sigma^2}\frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} \\
&\quad - \frac{\beta^2}{\sigma^2 + \beta^2}\frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2}.
\end{aligned}$$

Similarly, we have

$$T_{\text{model-X}}(\mathbf{Y}, \tilde{\mathbf{X}}_*, \mathbf{Z}_*) \mid \mathbf{Y}, \mathbf{Z}_* \sim \mathcal{N}\left(0, \sigma_0^2(\mathbf{Y}, \mathbf{Z}_*) = \left(1 - \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2}\right)\right),$$

and thus we reject if $T_{\text{model-X}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \geq z_{1-\alpha}\sqrt{\sigma_0^2(\mathbf{Y}, \mathbf{Z}_*)}$. Under $\beta = h/\sqrt{n}$, the power is

$$\begin{aligned}
&\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{model-X}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \geq z_{1-\alpha}\sqrt{\sigma_0^2(\mathbf{Y}, \mathbf{Z}_*)}\right) \\
&= \mathbb{E}\left[\mathbb{P}_{\beta=h/\sqrt{n}}\left(T_{\text{model-X}}(\mathbf{Y}, \mathbf{X}_*, \mathbf{Z}_*) \geq z_{1-\alpha}\sqrt{\sigma_0^2(\mathbf{Y}, \mathbf{Z}_*)} \mid \mathbf{Y}, \mathbf{Z}_*\right)\right] \\
&= \mathbb{E}\left[1 - \Phi\left(\frac{z_{1-\alpha}\sqrt{\sigma_0^2(\mathbf{Y}, \mathbf{Z}_*)} - \mu_{h/\sqrt{n}}(\mathbf{Y}, \mathbf{Z}_*)}{\sqrt{\sigma^2_{h/\sqrt{n}}(\mathbf{Y}, \mathbf{Z}_*)}}\right)\right].
\end{aligned} \tag{15}$$

Mean term. We first look at the term

$$\mu_\beta(\mathbf{Y}, \mathbf{Z}_*) = \frac{\beta}{\sigma^2 + \beta^2}\frac{\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta))}{\|\mathbf{Y}\|}.$$

Let $\epsilon = \mathbf{Y} - \mathbf{Z}(\theta + \beta\eta)$ be the residue vector that is independent of $\mathbf{Z}$.

$$\begin{aligned}
&\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta)) \\
&= \epsilon^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Z}(\theta + \beta\eta) + \epsilon) \\
&= \epsilon^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon + \epsilon^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Z}(\theta + \beta\eta) \\
&= \epsilon^\top\epsilon - \epsilon^\top(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon + \epsilon^\top(I - \mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Z}(\theta + \beta\eta).
\end{aligned}$$

We normalize the above expression by n and study each term. Note that we are under $\beta = h/\sqrt{n}$.

1. Since $\epsilon^\top\epsilon \sim (\sigma^2 + \beta^2)\chi^2_n$, $n^{-1}\epsilon^\top\epsilon \xrightarrow{p} \sigma^2$.
2. Let $A_i = (\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_i\mathbf{Z}_i^\top$, where $\mathbf{Z}_i$ is the ith row of $\mathbf{Z}_*$ as a column vector. Since all A_i's are exchangeable and $\sum_{i=1}^{n_*} A_i = I_p$, $\mathbb{E}[A_i] = I_p/n_*$.

$$\begin{aligned}
\mathbb{E}[n^{-1}\epsilon^\top(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon] &= \frac{\sigma^2 + \beta^2}{n}\mathbb{E}[\operatorname{tr}(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)] \\
&= \frac{\sigma^2 + \frac{1}{n}h^2}{n}\mathbb{E}[\operatorname{tr}((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}(\mathbf{Z}^\top\mathbf{Z})] \\
&= \frac{\sigma^2 + \frac{1}{n}h^2}{n}\operatorname{tr}(\mathbb{E}[(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}(\mathbf{Z}^\top\mathbf{Z})]) \\
&= \frac{\sigma^2 + \frac{1}{n}h^2}{n}\operatorname{tr}\left(\mathbb{E}\left[\sum_{j=1}^{n} A_j\right]\right) \\
&= \left(\sigma^2 + \frac{1}{n}h^2\right)\operatorname{tr}(I_p/n_*) \\
&= \left(\sigma^2 + \frac{1}{n}h^2\right)\frac{p}{n_*} \to \sigma^2\kappa\kappa_*.
\end{aligned}$$

Note that

$$\begin{aligned}
0 &= \operatorname{Var}\left(\operatorname{tr}(\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top)\right) \\
&= \operatorname{Var}\left(\operatorname{tr}\left(\sum_{i=1}^{n_*} A_i\right)\right) \\
&= n_*\operatorname{Var}(\operatorname{tr}(A_1)) + n_*(n_* - 1)\operatorname{Cov}(\operatorname{tr}(A_1), \operatorname{tr}(A_2)) \\
&\Rightarrow \operatorname{Cov}(\operatorname{tr}(A_1), \operatorname{tr}(A_2)) = -\frac{1}{n_* - 1}\operatorname{Var}(\operatorname{tr}(A_1)).
\end{aligned}$$

Thus, we have

$$\begin{aligned}\frac{\operatorname{Var}\left(\mathbb{E}[n^{-1}\epsilon^\top(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon \mid \mathbf{Z}_*]\right)}{\sigma^2+h^2/n} &= n^{-2}\operatorname{Var}\left(\operatorname{tr}(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\right)\\ &= n^{-2}\operatorname{Var}\left(\operatorname{tr}\left(\sum_{i=1}^n A_i\right)\right)\\ &= n^{-1}\operatorname{Var}(\operatorname{tr}(A_1)) + n^{-1}(n-1)\operatorname{Cov}(\operatorname{tr}(A_1),\operatorname{tr}(A_2))\\ &= n^{-1}\operatorname{Var}(\operatorname{tr}(A_1)) + n^{-1}(n-1)\left(-\frac{1}{n_*-1}\operatorname{Var}(\operatorname{tr}(A_1))\right)\\ &= \frac{n_*-n}{n(n_*-1)}\operatorname{Var}(\operatorname{tr}(A_1)) \le \frac{n_*-n}{n(n_*-1)} \to 0,\end{aligned}$$

where we use $\operatorname{tr}(A_1) = \lambda_{\max}(\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1) \le 1$. Too see this, note that $\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1$ is the $(1,1)$-entry of $\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top$, so $\lambda_{\max}(\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1) \le \lambda_{\max}(\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top) = 1$ (a projection matrix).

Recall the variance formula for Gaussian quadratic forms (Rencher and Schaalje 2008), i.e., if $W \sim \mathcal{N}(\mu,\Sigma)$, then

$$\operatorname{Var}(W^\top\Lambda W) = 2\operatorname{tr}(\Lambda\Sigma\Lambda\Sigma) + 4\mu^\top\Lambda\Sigma\Lambda\mu.$$

Thus, we have

$$\begin{aligned}\frac{\mathbb{E}\left[\operatorname{Var}\left(n^{-1}\epsilon^\top(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon \mid \mathbf{Z}_*\right)\right]}{(\sigma^2+h^2/n)^2} &= 2n^{-2}\mathbb{E}\left[\operatorname{tr}\left(\left(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\right)^2\right)\right]\\ &\le 2n^{-2}\mathbb{E}\left[\lambda_{\max}(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)^2\operatorname{tr}(I_p)\right]\\ &\le 2n^{-2}p \to 0.\end{aligned}$$

Here,

$$\lambda_{\max}(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top) \le \lambda_{\max}\left(\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top\right) = 1, \tag{16}$$

because $\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top$ is the matrix of the first n rows and n columns of $\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top$. To sum up,

$$\frac{1}{n}\epsilon^\top(\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon \xrightarrow{p} \sigma^2\kappa\kappa_*.$$

3. Trivially,

$$\mathbb{E}\left[\epsilon^\top(I-\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Z}(\theta+\beta\eta) \mid \mathbf{Z}_*\right] = 0.$$

As for the variance,

$$\begin{aligned}\frac{\operatorname{Var}\left[\epsilon^\top(I-\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Z}(\theta+\beta\eta) \mid \mathbf{Z}_*\right]}{\sigma^2+h^2/n} &= (\theta+\beta\eta)^\top\mathbf{Z}^\top(I-\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)^2\mathbf{Z}(\theta+\beta\eta)\\ &\le (\theta+\beta\eta)^\top\mathbf{Z}^\top\mathbf{Z}(\theta+\beta\eta).\end{aligned}$$

Then we have

$$\begin{aligned}&\mathbb{E}\left(\operatorname{Var}\left[\epsilon^\top(I-\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\mathbf{Z}(\theta+\beta\eta) \mid \mathbf{Z}_*\right]\right)\\ &\le n^{-2}(\sigma^2+h^2/n)(\theta+\beta\eta)^\top\mathbb{E}\left[\mathbf{Z}^\top\mathbf{Z}\right](\theta+\beta\eta)\\ &= \frac{\sigma^2+h^2/n}{n}(\theta+\beta\eta)^\top\Sigma_Z(\theta+\beta\eta) \to 0,\end{aligned}$$

using $\theta^\top \Sigma_Z \theta \to v_Z^2 < \infty$, $\beta = h/\sqrt{n}$, $\eta^\top \Sigma_Z \eta$ bounded and $\theta^\top \Sigma_Z \eta \le \sqrt{\theta^\top \Sigma_Z \theta \cdot \eta^\top \Sigma_Z \eta}$.

We have established

$$\frac{1}{n}\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta)) \xrightarrow{p} \sigma^2(1 - \kappa\kappa_*).$$

On the other hand, since

$$Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{h^2}{n} + (\theta + h\eta/\sqrt{n})^\top \Sigma_Z (\theta + h\eta/\sqrt{n}) + \sigma^2\right),$$

we can use (8) to get $\|\mathbf{Y}\|^2/n \xrightarrow{p} v_Z^2 + \sigma^2$. Now we can see that

$$\mu_{h/\sqrt{n}}(\mathbf{Y}, \mathbf{Z}_*) = \frac{h}{\sigma^2 + \frac{1}{n}h^2} \frac{\frac{1}{n}\mathbf{Y}^\top (I - \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)(\mathbf{Y} - \mathbf{Z}(\theta + \beta\eta))}{\frac{1}{\sqrt{n}}\|\mathbf{Y}\|} \xrightarrow{p} \frac{h(1 - \kappa\kappa_*)}{\sqrt{v_Z^2 + \sigma^2}}.$$

Variance term. Next, we look at

$$\sigma_\beta^2(\mathbf{Y}, \mathbf{Z}_*) = \frac{\sigma^2}{\sigma^2 + \beta^2} + \frac{\beta^2 - \sigma^2}{\beta^2 + \sigma^2} \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} - \frac{\beta^2}{\sigma^2 + \beta^2} \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2}.$$

We first note that

$$\frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} \le \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} \le \lambda_{\max}\left(\mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top\right) \le 1$$

by (16), so the last term in the expression of $\sigma_\beta^2(\mathbf{Y}, \mathbf{Z}_*)$ (recall $\beta = h/\sqrt{n}$) satisfies

$$\frac{\beta^2}{\sigma^2 + \beta^2} \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} \to 0.$$

Similarly,

$$\sigma_\beta^2(\mathbf{Y}, \mathbf{Z}_*) - \sigma_0^2(\mathbf{Y}, \mathbf{Z}_*) = \frac{\beta^2}{\sigma^2 + \beta^2}\left(2\frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2} - \frac{\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}}{\|\mathbf{Y}\|^2}\right) \to 0. \tag{17}$$

Next, we analyze $\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y}$.

$$\begin{aligned}&\mathbf{Y}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Y} \\ &= \epsilon^\top (\mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon + 2\epsilon^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta + \beta\eta) + (\theta + \beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta + \beta\eta).\end{aligned}$$

We again look at these terms one by one (after normalization by n).

1. We have already shown
$$\frac{1}{n}\epsilon^\top (\mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top)\epsilon \xrightarrow{p} \sigma^2 \kappa\kappa_*.$$

2.

$$\mathbb{E}\left[\epsilon^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \mid \mathbf{Z}_*\right] = 0.$$

$$\begin{aligned}
&\frac{\mathbb{E}\left(\mathrm{Var}\left[n^{-1}\epsilon^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \mid \mathbf{Z}_*\right]\right)}{\sigma^2 + h^2/n} \\
&= \frac{1}{n^2}\mathbb{E}\left((\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta)\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left((\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}_*^\top \mathbf{Z}_*(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta)\right) \\
&= \frac{1}{n^2}\mathbb{E}\left((\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta)\right) \\
&\leq \frac{1}{n^2}\mathbb{E}\left((\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta)\right) \\
&= \frac{1}{n}(\theta+\beta\eta)^\top \Sigma_Z(\theta+\beta\eta) \to 0,
\end{aligned}$$

where the second to last step is because $\lambda_{\max}(\mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top) \leq 1$ and the last step is because $\theta^\top \Sigma_Z \theta \to v_Z^2 < \infty$, $\beta = h/\sqrt{n}$, $\eta^\top \Sigma_Z \eta$ bounded and $\theta^\top \Sigma_Z \eta \leq \sqrt{\theta^\top \Sigma_Z \theta \cdot \eta^\top \Sigma_Z \eta}$.

3. We now assume without loss of generality that $\Sigma_Z = I_p$, which we can achieve by absorbing $\Sigma_Z^{1/2}$ into $(\theta+\beta\eta)$. We have the loose bounds

$$\frac{1}{n}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \geq 0,$$

$$\begin{aligned}
&\frac{1}{n}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \\
&= \frac{1}{nn_*}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}\left(\frac{\mathbf{Z}_*^\top \mathbf{Z}_*}{n_*}\right)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \\
&\leq \frac{1}{\lambda_{\min}\left(\frac{\mathbf{Z}_*^\top \mathbf{Z}_*}{n_*}\right)}\frac{1}{nn_*}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \\
&\leq \frac{\lambda_{\max}\left(\frac{\mathbf{Z}\mathbf{Z}^\top}{p}\right)}{\lambda_{\min}\left(\frac{\mathbf{Z}_*^\top \mathbf{Z}_*}{n_*}\right)}\frac{p}{nn_*}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \\
&= \frac{\lambda_{\max}\left(\frac{\mathbf{Z}\mathbf{Z}^\top}{p}\right)}{\lambda_{\min}\left(\frac{\mathbf{Z}_*^\top \mathbf{Z}_*}{n_*}\right)}\frac{p}{n_*}(\theta+\beta\eta)^\top \frac{\mathbf{Z}^\top \mathbf{Z}}{n}(\theta+\beta\eta) \\
&\xrightarrow{p} \frac{(1+\sqrt{1/\kappa})^2}{(1-\sqrt{\kappa\kappa_*})^2}\kappa\kappa_* v_Z^2,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{n}(\theta+\beta\eta)^\top \mathbf{Z}^\top \mathbf{Z}(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\mathbf{Z}^\top \mathbf{Z}(\theta+\beta\eta) \\
&\leq (\theta+\beta\eta)^\top \frac{\mathbf{Z}^\top \mathbf{Z}}{n}(\theta+\beta\eta) \\
&\xrightarrow{p} v_Z^2,
\end{aligned}$$

Thus, we already have

$$\mathbb{P}_{\beta=h/\sqrt{n}}\left(1-\frac{\sigma^2\kappa\kappa_*+v_Z^2\min\left(1,\kappa\kappa_*\frac{(1+\sqrt{1/\kappa})^2}{(1-\sqrt{\kappa\kappa_*})^2}\right)}{\sigma^2+v_Z^2}\le\sigma_\beta^2(\mathbf{Y},\mathbf{Z}_*)\le 1-\frac{\sigma^2\kappa\kappa_*}{\sigma^2+v_Z^2}\right)\to 1.$$

Since both the lower and upper bounds in the above equation are positive, together with (17) we get

$$\frac{\sigma_0^2(\mathbf{Y},\mathbf{Z}_*)}{\sigma_\beta^2(\mathbf{Y},\mathbf{Z}_*)}\xrightarrow{p}1.$$

Then we get

$$\limsup(15)\le\Phi\left(\frac{h\tau\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2\max\left(0,\frac{1-\frac{(1+\sqrt{1/\kappa})^2}{(1-\sqrt{\kappa\kappa_*})^2}\kappa\kappa_*}{1-\kappa\kappa_*}\right)}}-z_{1-\alpha}\right),$$

$$\liminf(15)\ge\Phi\left(\frac{h\tau\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2\frac{1}{1-\kappa\kappa_*}}}-z_{1-\alpha}\right).$$

□

Conjecture 1. *In Setting 1, if there are m additional data points $(X_i,Z_i)_{i=n+1}^{n+m}$, $n_*=n+m$, $n/n_*\to\kappa_*$ and $\kappa\kappa_*<1$, then the conditional CRT with statistic T_{MC} has asymptotic power equal to that of a z-test with standardized effect size*

$$\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2(1-\kappa_*)}}.$$

Analysis of Conjecture 1. Finding the asymptotic power is finding the exact limit of (15), which requires a more careful analysis. Picking up from the end of the proof of Theorem 4, we now find the limit of the expectation of the third term in the variance decomposition normalized by n, i.e.,

$$\lim\mathbb{E}\left[\frac{1}{n}(\theta+\beta\eta)^\top\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}(\theta+\beta\eta)\right].$$

Recall that we are assuming $\Sigma_Z=I$ without loss of generality. We will show $\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}]=c(p,n,n_*)I_p$ is a mutliple of I_p. Once this is done, we will find the limit of $c(p,n,n_*)/n$.

To see this, we show that the expectation (a $p\times p$ matrix) is invariant under orthogonal transformation. Let Q be any $p\times p$ orthogonal matrix.

$$\begin{aligned}\mathbb{E}[Q\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}Q^\top]&=\mathbb{E}[Q\mathbf{Z}^\top\mathbf{Z}Q^\top(Q\mathbf{Z}_*^\top\mathbf{Z}_*Q^\top)^{-1}Q\mathbf{Z}^\top\mathbf{Z}Q^\top]\\&=\mathbb{E}[\mathbf{W}^\top\mathbf{W}(\mathbf{W}_*^\top\mathbf{W}_*)^{-1}\mathbf{W}^\top\mathbf{W}]\ (\mathbf{W}_*=\mathbf{Z}_*Q^\top\overset{d}{=}\mathbf{Z}_*)\\&=\mathbb{E}[\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}].\end{aligned}$$

This shows the expectation must be a multiple of I_p. Now we consider

$$\begin{aligned}
n_* I_p &= \mathbb{E}[\mathbf{Z}_*^\top \mathbf{Z}_*] \\
&= \mathbb{E}[\mathbf{Z}_*^\top \mathbf{Z}_* (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_*^\top \mathbf{Z}_*] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n_*} \mathbf{Z}_i \mathbf{Z}_i^\top\right)(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\left(\sum_{i=1}^{n_*} \mathbf{Z}_i \mathbf{Z}_i^\top\right)\right] \\
&= \sum_{i=1}^{n_*} \mathbb{E}\left[\mathbf{Z}_i \mathbf{Z}_i^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_i \mathbf{Z}_i^\top\right] + \sum_{i \neq j} \mathbb{E}\left[\mathbf{Z}_i \mathbf{Z}_i^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_j \mathbf{Z}_j^\top\right] \\
&= n_* \mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \mathbf{Z}_1^\top] + n_*(n_*-1)\mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_2 \mathbf{Z}_2^\top] \\
&= n_* a_{p,n_*} I_p + n_*(n_*-1) b_{p,n_*} I_p,
\end{aligned}$$

where $\mathbf{Z}_i$ is the ith row of $\mathbf{Z}_*$ as a column vector and

$$\begin{aligned}
\mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \mathbf{Z}_1^\top] &= a_{p,n_*} I_p, \\
\mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_2 \mathbf{Z}_2^\top] &= b_{p,n_*} I_p
\end{aligned}$$

are multiples of I_p by the same argument. From this we get $a_{p,n_*} + (n_*-1) b_{p,n_*} = 1$. Similarly,

$$\begin{aligned}
&\mathbb{E}[\mathbf{Z}^\top \mathbf{Z} (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}^\top \mathbf{Z}] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^\top\right)(\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1}\left(\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i^\top\right)\right] \\
&= n\mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \mathbf{Z}_1^\top] + n(n-1)\mathbb{E}[\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_2 \mathbf{Z}_2^\top] \\
&= n a_{p,n_*} I_p + n(n-1) b_{p,n_*} I_p \\
&= n a_{p,n_*} I_p + n(n-1)\frac{1 - a_{p,n_*}}{n_* - 1} I_p.
\end{aligned}$$

Therefore, by representing b_{p,n_*} with a_{p,n_*}, we have

$$\begin{aligned}
n_*^{-1} c(p,n,n_*) I_p &= \mathbb{E}[\mathbf{Z}^\top \mathbf{Z} (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}^\top \mathbf{Z}/n_*] \\
&= \frac{n}{n_*} a_{p,n_*} I_p + n(n-1)\frac{1 - a_{p,n_*}}{n_*(n_*-1)} I_p \\
&= \frac{n}{n_*}\left(\left(1 - \frac{n-1}{n_*-1}\right) a_{p,n_*} + \frac{n-1}{n_*-1}\right) I_p.
\end{aligned}$$

We have shown that

$$n_*^{-1} c(p,n,n_*) = \frac{n}{n_*}\left(\left(1 - \frac{n-1}{n_*-1}\right) a_{p,n_*} + \frac{n-1}{n_*-1}\right). \tag{18}$$

Recall that we are interested in the limit of this term, so we can focus on the limit of a_{p,n_*}.

$$\begin{aligned}
a_{p,n_*} &= \mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{Z}_1 \mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \mathbf{Z}_1^\top}{p}\right)\right] \\
&= \mathbb{E}\left[\operatorname{tr}\left(\frac{\mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \mathbf{Z}_1^\top \mathbf{Z}_1}{p}\right)\right] \\
&= \mathbb{E}\left[\mathbf{Z}_1^\top (\mathbf{Z}_*^\top \mathbf{Z}_*)^{-1} \mathbf{Z}_1 \cdot \frac{\|\mathbf{Z}_1\|^2}{p}\right].
\end{aligned}$$

To study this expectation, note that

$$\begin{aligned}\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\right] &= \mathbb{E}\left[\mathrm{tr}\left(\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\mathbf{Z}_1^\top\right)\right].\end{aligned}$$

Note that by symmetry,

$$\begin{aligned}n_*\mathbb{E}\left[\mathrm{tr}\left((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\mathbf{Z}_1^\top\right)\right] &= \sum_{i=1}^{n_*}\mathbb{E}\left[\mathrm{tr}\left((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_i\mathbf{Z}_i^\top\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left(\sum_{i=1}^{n_*}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_i\mathbf{Z}_i^\top\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\left(\sum_{i=1}^{n_*}\mathbf{Z}_i\mathbf{Z}_i^\top\right)\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left((\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\left(\mathbf{Z}_*^\top\mathbf{Z}_*\right)\right)\right] \\ &= \mathbb{E}\left[\mathrm{tr}\left(I_p\right)\right] = p \Rightarrow \mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\right] = p/n_*.\end{aligned}$$

We can also note that $\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1$ is the first diagonal element of the projection matrix $\mathbf{Z}_*(\mathbf{Z}_*^\top\mathbf{Z}_*)\mathbf{Z}_*^\top$, whose eigenvalues are all 0 or 1. Thus, $0 \le \mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1 \le 1$.

Note that all the equations above are exact and no limit has been taken yet. Now we show the number sequence $a_{p,n_*} \to \kappa\kappa_*$, which we recall is the limit of p/n_*.

For any $\delta > 0$, note that $\|\mathbf{Z}_1\|^2 \sim \chi_p^2$.

$$\begin{aligned}&\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\right] \\ &= \mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}\le 1+\delta)\right] + \mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta)\right] \\ &\le (1+\delta)\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\right] + \mathbb{E}\left[1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta)\right] \\ &= (1+\delta)\frac{p}{n_*} + \mathbb{E}\left[\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta)\right] \\ &\le (1+\delta)\frac{p}{n_*} + \sqrt{\mathbb{E}\left[\frac{\|\mathbf{Z}_1\|^4}{p^2}\right]\mathbb{E}\left[\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta)^2\right]} \\ &= (1+\delta)\frac{p}{n_*} + \sqrt{\frac{p^2+2p}{p^2}\mathbb{P}\left(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta\right)} \\ &= (1+\delta)\frac{p}{n_*} + \sqrt{1+\frac{2}{p}}\sqrt{\mathbb{P}\left(\frac{\|\mathbf{Z}_1\|^2}{p}> 1+\delta\right)} \\ &\le (1+\delta)\frac{p}{n_*} + \sqrt{1+\frac{2}{p}}\sqrt{\mathbb{P}\left(\left|\frac{\|\mathbf{Z}_1\|^2}{p}-1\right|> \delta\right)} \\ &\le (1+\delta)\frac{p}{n_*} + \sqrt{1+\frac{2}{p}}\sqrt{\frac{\mathrm{Var}\left(\|\mathbf{Z}_1\|^2/p\right)}{\delta^2}} \\ &= (1+\delta)\frac{p}{n_*} + \sqrt{1+\frac{2}{p}}\sqrt{\frac{2}{p\delta^2}} \to (1+\delta)\kappa\kappa_*.\end{aligned}$$

Since δ can be arbitrarily small, this shows

$$\limsup \mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\right]\le\kappa\kappa_*. \tag{19}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\right]\\
&=\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}\ge1-\delta)\right]+\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}<1-\delta)\right]\\
&\ge\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}\ge1-\delta)\right]\\
&\ge(1-\delta)\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}\ge1-\delta)\right]\\
&=(1-\delta)\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\right]-(1-\delta)\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}<1-\delta)\right]\\
&=(1-\delta)\frac{p}{n_*}-(1-\delta)\mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}<1-\delta)\right]\\
&\ge(1-\delta)\frac{p}{n_*}-(1-\delta)\mathbb{E}\left[1\cdot\mathbb{I}(\frac{\|\mathbf{Z}_1\|^2}{p}<1-\delta)\right]\\
&\ge(1-\delta)\frac{p}{n_*}-(1-\delta)\mathbb{P}\left(\left|\frac{\|\mathbf{Z}_1\|^2}{p}-1\right|>\delta\right)\\
&\ge(1-\delta)\frac{p}{n_*}-(1-\delta)\frac{\mathrm{Var}\left(\|\mathbf{Z}_1\|^2/p\right)}{\delta^2}\\
&=(1-\delta)\frac{p}{n_*}-(1-\delta)\frac{2}{p\delta^2}\to(1-\delta)\kappa\kappa_*.
\end{aligned}$$

Since δ can be arbitrarily small, this shows

$$\liminf \mathbb{E}\left[\mathbf{Z}_1^\top(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}_1\cdot\frac{\|\mathbf{Z}_1\|^2}{p}\right]\ge\kappa\kappa_*. \tag{20}$$

Equations (19) and (20) show $a_{p,n_*}\to\kappa\kappa_*$. This together with equation (18) shows

$$\begin{aligned}
n^{-1}c(p,n,n_*)&=\left(1-\frac{n-1}{n_*-1}\right)a_{p,n_*}+\frac{n-1}{n_*-1}\\
&\to(1-\kappa_*)\,\kappa\kappa_*+\kappa_*\\
&=\kappa_*(1+\kappa(1-\kappa_*)).
\end{aligned}$$

To sum up, we have shown

$$\frac{1}{n}\mathbb{E}\left[(\theta+\beta\eta)^\top\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}(\theta+\beta\eta)\right]=\frac{c(p,n,n_*)}{n}(\theta+\beta\eta)^\top\Sigma_Z(\theta+\beta\eta)\to\kappa_*(1+\kappa(1-\kappa_*))v_Z^2.$$

We conjecture that actually (e.g., if we can show the variance converges to 0)

$$\frac{1}{n}(\theta+\beta\eta)^\top\mathbf{Z}^\top\mathbf{Z}(\mathbf{Z}_*^\top\mathbf{Z}_*)^{-1}\mathbf{Z}^\top\mathbf{Z}(\theta+\beta\eta)\xrightarrow{p}\kappa_*(1+\kappa(1-\kappa_*))v_Z^2,$$

in which case (15) converges to

$$\Phi\left(\frac{h\sqrt{1-\kappa\kappa_*}}{\sqrt{\sigma^2+v_Z^2(1-\kappa_*)}}-z_{1-\alpha}\right).$$

□

Theorem 10. *Let J_0 and J_1 form a partition of $\{1,2,\ldots,p\}$ and $|J_0|/p \to \gamma \in (0,1)$. Consider p random variables $p_j = 1 - F_j(T_j)$, which should be thought of as p-values. Assume the following conditions.*

(a) For $j \in J_0$, $1 - F_j(T_j) \sim \text{Unif}[0,1]$; for any $t \in \mathbb{R}$, $F_j(t) \xrightarrow{p} F^{(0)}(t)$ and for $j \in J_1$, $F_j(t) \xrightarrow{p} F^{(1)}(t)$. $F^{(0)}$ and $F^{(1)}$ are deterministic CDFs of random variables with common connected support and continuous densities on their support. For $j' \in J_1$, $T_{j'} \xrightarrow{d} T^{(1)}$, which has the same support as $F^{(0)}$ and $F^{(1)}$ and continuous density on the support.

(b) Within J_0 or J_1, p_j's are exchangeable.

(c) For distinct $j_1, j_2 \in J_0$ and distinct $j_3, j_4 \in J_1$, the following two pairs of random variables are asymptotically pairwise independent: (T_{j_1}, T_{j_2}) and (T_{j_3}, T_{j_4}). That is, both pairs converge in distribution to a bivariate random vector (not necessarily the same random vector) with independent components.

Then let G be the CDF of $1 - F^{(1)}(T^{(1)})$, $q \in (0,1)$, and

$$t(g) = \max\{t \in (0,1] : g(t) \le q\}.$$

When the set is empty, define $t(g) = 0$. Let

$$g_{\text{BH}}(t) = \frac{t}{\gamma t + (1-\gamma)G(t)}$$

and

$$g_{\text{AdaPT}}(t) = \frac{\gamma t + (1-\gamma)\left(1 - G(1-t)\right)}{\gamma t + (1-\gamma)G(t)}.$$

The for $t_{\text{BH}} = t(g_{\text{BH}})$ and almost every $q \in (0,1)$, at least one of the following cases is true: (i) $t_{\text{BH}} \in (0,1)$ and $g'_{BH}(t_{\text{BH}}) \neq 0$, (ii) $t_{\text{BH}} = 1$ and $g_{\text{BH}}(1) < q$, or (iii) $t_{BH} = 0$. In cases (i) or (ii), for the BH procedure at level q, the FDP and realized power converge in probability to

$$\frac{\gamma t_{\text{BH}}}{\gamma t_{\text{BH}} + (1-\gamma)G(t_{\text{BH}}))} \quad \textit{and} \quad G(t_{\text{BH}}),$$

respectively. In case (i), the asymptotic realized power simplifies to γq. In case (iii), the realized power converges in probability to 0.

For $t_{\text{AdaPT}} = t(g_{\text{AdaPT}})$ and almost every $q \in (0,1)$, at least one of the following cases is true: (i) $t_{\text{AdaPT}} \in (0,1)$ and $g'_{\text{AdaPT}}(t_{\text{AdaPT}}) \neq 0$, (ii) $t_{\text{AdaPT}} = 1$ and $g_{\text{AdaPT}}(1) < q$, or (iii) $t_{\text{AdaPT}} = 0$. In cases (i) or (ii), for the AdaPT procedure at level q, the FDP and realized power converge in probability to

$$\frac{\gamma t_{\text{AdaPT}}}{\gamma t_{\text{AdaPT}} + (1-\gamma)G(t_{\text{AdaPT}}))} \quad \textit{and} \quad G(t_{\text{AdaPT}}),$$

respectively. In case (iii), the realized power converges in probability to 0.

Proof of Theorem 10. We only prove the case for AdaPT, because the proof for BH is similar and slightly easier. We suppress the subsript in g_{AdaPT}.

If $q > g(1)$, then (ii) holds. If $q < \inf\{g(t) : t \in (0,1]\}$, then (iii) holds. If $q \in (\inf\{g(t) : t \in (0,1], g(1))$, then because g is continuous, we can see that $t_{\text{AdaPT}} \in (0,1)$ (note that we are considering maximum and $(0,1]$ is closed on the right, so the maximum must exist and not equal to 1). And in this case, $g(t_{\text{AdaPT}}) = q$, since otherwise $g(t_{\text{AdaPT}}) < q$ and t_{AdaPT} could be smaller because of g's continuity. Next we show for almost every $q \in (\inf\{g(t) : t \in (0,1], g(1))$, case (i) holds. We just need to show that the set $\{g(t) : g'(t) = 0\}$ has measure zero, which is a simple application of Sard's theorem (take $n = m = k = 1$, $f(x) = g(\arctan(x)/\pi + 1/2)$, where $\arctan(x)/\pi + 1/2$ is just some function that maps $\mathbb{R}$ to $(0,1)$ with continuous nonzero derivative), where g' is continuous on (0,1) because $F^{(1)}$, $T^{(1)}$ and thus G have continuous densities on a common connected support.

Lemma 7 (Sard's theorem, Sard (1942))**.** *Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be k times continuously differentiable, where $k \geq \max(n - m + 1, 1)$. Let A be the set of points $x \in \mathbb{R}^n$ such that the Jacobian matrix of f has rank smaller than m. Then the image $f(A)$ has Lebesgue measure zero in $\mathbb{R}^m$.*

In case (i), $t_{\text{AdaPT}} \in (0,1)$, we have established that $g(t_{\text{AdaPT}}) = q$. Since $g'(t_{\text{AdaPT}}) \neq 0$, we must have $g'(t_{\text{AdaPT}}) < 0$, since otherwise t_{AdaPT} could also be smaller. Thus, in case (i), $g'(t_{\text{AdaPT}}) < 0$ and for any sufficiently small $\varepsilon > 0$ there exists a point t^* in $(t_{\text{AdaPT}} - \varepsilon, t_{\text{AdaPT}})$ such that $g(t^*) < q$. In case (ii), since g is continuous and $g(1) < q$, it also holds that for any sufficiently small $\varepsilon > 0$ there exists a point t^* in $(t_{\text{AdaPT}} - \varepsilon, t_{\text{AdaPT}})$ such that $g(t^*) < q$.

Let

$$\hat{t}_p = \min\{t \in (0,1] : g_p(t) \equiv \frac{1/p + \#\{j : p_j \geq 1 - t\}/p}{\#\{j : p_j \leq t\}/p} \leq q\}.^7$$

We analyze cases (i) and (ii). We begin by showing $\hat{t}_p \xrightarrow{p} t_{\text{AdaPT}}$. Take any sufficiently small $\varepsilon > 0$. We have established that there exists a point t^* in $(t_{\text{AdaPT}} - \varepsilon, t_{\text{AdaPT}})$ such that $g(t^*) < q$. Then

$$\begin{aligned}
\mathbb{P}(\hat{t}_p > t_{\text{AdaPT}} - \varepsilon) &\geq \mathbb{P}(\hat{t}_p \geq t^*) \\
&\geq \mathbb{P}(g_p(t^*) \leq q) \\
&\geq \mathbb{P}(|g_p(t^*) - g(t^*)| < |g(t^*) - q|) \to 1.
\end{aligned}$$

In case (ii), we get $\mathbb{P}(\hat{t}_p \leq t_{\text{AdaPT}} = 1) = 1$ for free and the proof is concluded. Next we consider case (i) and assume $\varepsilon \in (0, 1 - t_{\text{AdaPT}})$. Choose $\delta_1 \in (0, \min(1 - t_{\text{AdaPT}} - \varepsilon, 1 - G(t_{\text{AdaPT}} + \varepsilon)))$. Let $\delta_3 = \min\{g(t) - q : t \leq [t_{\text{AdaPT}} + \varepsilon, 1]\}$, which is positive since otherwise g could attain a value no more than q in $[t_{\text{AdaPT}} + \varepsilon, 1]$, violating t_{AdaPT}'s definition. Now observe that the function $\frac{\gamma x + (1-\gamma)y}{\gamma z + (1-\gamma)w}$ is continuous in (x, y, z, w) on $\{(x, y, z, w) \in [0,1]^4 : z, w \geq \min(1 - t_{\text{AdaPT}} - \varepsilon, 1 - G(t_{\text{AdaPT}} + \varepsilon) - \varepsilon) - \delta_1\}$, and thus uniformly continuous. So we can choose $\delta_2 \in (0, \delta_1)$ such that whenever $(x, y, z, w), (x', y', z', w') \in \{(x, y, z, w) \in [0,1]^4 : z, w \geq \min(1 - t_{\text{AdaPT}} - \varepsilon, 1 - G(t_{\text{AdaPT}} + \varepsilon))\}$ and $|x - x'|, |y - y'|, |z - z'|, |w - w'| < \delta_2$, $|\frac{\gamma x + (1-\gamma)y}{\gamma z + (1-\gamma)w} - \frac{\gamma x' + (1-\gamma)y'}{\gamma z' + (1-\gamma)w'}| < \delta_3$.

Now we show the empirical CDFs of the non-null and null p-values converge pointwise, which implies uniform convergence by Lemma 8. Let $\hat{G}_p^{(1)}(t) = \#\{j \in J_1 : p_j \leq t\}/|J_1|$. Due to exchangeability, for any $j \in J_1$,

$$\begin{aligned}
\mathbb{E}[\hat{G}_p^{(1)}(t)] &= \mathbb{P}[p_j \leq t] = \mathbb{P}(1 - F_j(T_j) \leq t) \\
&\to \mathbb{P}(1 - F^{(1)}(T^{(1)}) \leq t) = G_1(t).
\end{aligned}$$

[7] The term $1/p$ does not matter asymptotically, in that we can still consider the numerator as an empirical CDF of the p_j's.

As for the variance, we have for any distinct $j, k \in J_1$,

$$\mathrm{Var}[\hat{G}_p^{(1)}(t)] = \frac{1}{(1-\gamma)p}\mathrm{Var}[\mathbf{1}(p_j \le t)] + \frac{(1-\gamma)p((1-\gamma)p-1)}{(1-\gamma)^2p^2}\mathrm{Cov}(\mathbf{1}(p_j \le t), \mathbf{1}(p_k \le t)).$$

Since all the F_j's converges in distribution to deterministic and continuous CDFs, by the asymptotic pairwise independence, $\mathrm{Var}[\hat{G}_p^{(1)}(t)] \to 0$, therefore establishing $\hat{G}_p^{(1)}(t) \xrightarrow{p} G_1(t)$. We can show the empirical CDF of the null p-values converges in probability in the same way.

By Lemma 8, with probability converging to 1, all the CDFs fall within a δ_2-neighborhood around the limit CDFs, and by the previously showed uniform continuity, $|g_p(t) - g(t)| < \delta_3$ for $t \in [t_{\mathrm{AdaPT}} + \varepsilon, 1]$. By the definition of δ_3, this means with probability converging to 1, $g_p(t) > q$ for all $t \in [t_{\mathrm{AdaPT}} + \varepsilon, 1]$. Now we have

$$\mathbb{P}(\hat{t}_p < t_{\mathrm{AdaPT}} + \varepsilon) = \mathbb{P}(g_p(t) > q, \forall t \in [t_{\mathrm{AdaPT}} + \varepsilon, 1]) \to 1.$$

Combining the results, we have

$$\lim_{p\to\infty} \mathbb{P}(|\hat{t}_p - t_{\mathrm{AdaPT}}| < \varepsilon) = 1.$$

Due to the uniform convergence of the CDFs, the results on the expressions on the asymptotic FDP and realized power then follow by noticing that the FDP and the realized power are

$$\frac{\#\{j \in J_0 : p_j \le \hat{t}_p\}}{\#\{j : p_j \le \hat{t}_p\}} \quad \text{and} \quad \frac{\#\{j \in J_1 : p_j \le \hat{t}_p\}}{|J_1|},$$

respectively.

Finally, we consider case (iii). In this case, we have $q < \inf\{g(t) : t \in (0, 1]\}$. Take any small positive δ, we have $q < \inf\{g(t) : t \in [\delta, 1]\}$. Since 0 is excluded from $[\delta, 1]$, we can again use the uniform continuity argument we used before, where we consider $g_p(t)$ as a function of four inputs, and the two inputs in the denominator are bounded away from 0. In this way, we can show $g_p(t) \xrightarrow{p} t$ uniformly for $t \in [\delta, 1]$, so that we have

$$\mathbb{P}(\hat{t}_p < \delta) = \mathbb{P}(g_p(t) \ge q, \forall t \in [\delta, 1]) \to 1.$$

Thus, the asymptotic realized power satisfies

$$\begin{aligned}
\frac{\#\{j \in J_1 : p_j \le \hat{t}_p\}}{|J_1|} &= \mathbf{1}_{\{\hat{t}_p < \delta\}} \frac{\#\{j \in J_1 : p_j \le \hat{t}_p\}}{|J_1|} + \mathbf{1}_{\{\hat{t}_p \ge \delta\}} \frac{\#\{j \in J_1 : p_j \le \hat{t}_p\}}{|J_1|} \\
&\le \underbrace{\mathbf{1}_{\{\hat{t}_p < \delta\}}}_{\xrightarrow{p} 1} \underbrace{\frac{\#\{j \in J_1 : p_j \le \delta\}}{|J_1|}}_{\xrightarrow{p} G(\delta)} + \underbrace{\mathbf{1}_{\{\hat{t}_p \ge \delta\}}}_{\xrightarrow{p} 0} \underbrace{\frac{\#\{j \in J_1 : p_j \le \hat{t}_p\}}{|J_1|}}_{\le 1} \\
&\xrightarrow{p} G(\delta).
\end{aligned}$$

Since δ can be arbitrarily small and G has no point mass at 0 because all distributions considered here are continuous, we have shown the realized power on the left hand side converges in distribution to 0.

□

Lemma 8. *Let $\{G_n\}$ be a sequence of random CDFs. If for every $t \in \mathbb{R}$, $G_n(t) \xrightarrow{p} G(t)$, with G being a continuous CDF, then the convergence is uniform in t, in the sense that*

$$\sup_{t\in\mathbb{R}} |G_n(t) - G(t)| \xrightarrow{p} 0.$$

Proof of Lemma 8. Take any $\varepsilon > 0$. Find a finite number of points $x_1, x_2, \ldots, x_N$ such that $G(x_1) \le \varepsilon/2$, $G(x_i) - G(x_{i-1}) \le \varepsilon/2$ and $1 - G(x_N) \le \varepsilon/2$. We have

$$\mathbb{P}\left(\max_{1\le i\le N} |G_n(x_i) - G(x_i)| \le \varepsilon/2\right) \to 1.$$

Now for any $x \in [x_{i-1}, x_i]$ ($x_0 = -\infty, x_{N+1} = \infty$), on the event $\max_{1\le i\le N} |G_n(x_i) - G(x_i)| \le \varepsilon/2$,

$$G_n(x) \ge G_n(x_{i-1}) \ge G(x_{i-1}) - \varepsilon/2 \ge G(x_i) - \varepsilon/2 - \varepsilon/2 \ge G(x) - \varepsilon,$$

$$G_n(x) \le G_n(x_i) \le G(x_i) + \varepsilon/2 \le G(x_{i-1}) + \varepsilon/2 + \varepsilon/2 \le G(x) + \varepsilon.$$

Thus,

$$\mathbb{P}(\sup_t |G_n(t) - G(t)| < \varepsilon) \ge \mathbb{P}(\max_{1\le i\le N} |G_n(x_i) - G(x_i)| \le \varepsilon/2) \to 1.$$

□

Theorem 5. *In Setting 2, for Lebesgue-almost-every $q \in (0, 1)$, BH or AdaPT at level q using CRT p-values based on the statistics in Section 2.2 (respectively, their absolute values) have the following one-sided (respectively, two-sided) effective π_μ's with respect to BH or AdaPT at level q:*

1. *For the marginal covariance statistic, the effective π_μ is the distribution of $\frac{1}{\sqrt{\sigma^2+\kappa\mathbb{E}[B_0^2]}} B_0$.*
2. *For the OLS statistic, assuming $\kappa < 1$, the effective π_μ is the distribution of $\frac{\sqrt{1-\kappa}}{\sigma} B_0$.*
3. *For the distilled lasso statistic, the effective π_μ is the distribution of $\frac{1}{\tau_\lambda} B_0$.*

Proof of Theorem 5. We prove the case of one-sided p-values, while it is clear that the heart of these results is at the asymptotic pairwise independence of the variable important statistics T_j's, and switching to two-sided p-values (or other reasonable p-values) only effectively changes using T_j into using $|T_j|$, the results for which can be established almost identically. We use Φ_{a^2} to denote the CDF of $\mathcal{N}(0, a^2)$.

1. *Marginal covariance.* Let $T_j = n^{-1/2}\mathbf{Y}^\top \mathbf{X}_j$ and F_j be the CDF of $\mathcal{N}(0, \|\mathbf{Y}\|_2^2/n)$. Now we check the conditions of Theorem 10.
   (a) It is clear that $1 - F_j(T_j) \sim \text{Unif}[0, 1]$ for the null variables. Note that $\|\mathbf{Y}\|^2/n \xrightarrow{p} \sigma^2 + \kappa\mathbb{E}[B_0^2]$ (see, e.g., the proof of Theorem 1), so for any j and t, $F_j(t) \xrightarrow{p} \Phi_{\sigma^2+\kappa\mathbb{E}[B_0^2]}(t)$. We will find $\mathcal{L}(T^{(1)})$ with Lemma 9.
   (b) This is true because β_j's are i.i.d.
   (c) We verify this condition by introducing Lemma 9, which also completes part (a).

   Lemma 9. *In Setting 2, for distinct j and k we have*

$$\begin{pmatrix} T_j \\ T_k \end{pmatrix} - \begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma^2 + \kappa\mathbb{E}[B_0^2] & 0 \\ 0 & \sigma^2 + \kappa\mathbb{E}[B_0^2] \end{bmatrix}\right).$$

To sum up, the p-values we obtain from marginal covariance satisfy the conditions of Theorem 10 with $F^{(1)} = \Phi_{\sigma^2+\kappa\mathbb{E}[B_0^2]}$ and $T^{(1)} \sim B_0 + \sqrt{\sigma^2 + \kappa\mathbb{E}[B_0^2]}W$, where W is a standard Gaussian random variable independent of B_0. We obtain the effective π_μ by dividing the T_j's by $\sqrt{\sigma^2 + \kappa\mathbb{E}[B_0^2]}$.

2. *OLS.* Let $T_j = \sqrt{n}\hat{\beta}_j$ be the (normalized) OLS estimate for covariate X_j. Let $\mathbf{X}_{-j}$ be $\mathbf{X}$ with the jth column removed and F_j be the CDF of $\mathcal{L}(T_j \,|\, \mathbf{Y}, \mathbf{X}_{-j})$ under the null. We now check the conditions of Theorem 10.

   (a) It is clear that $1 - F_j(T_j) \sim \text{Unif}[0,1]$ for the null variables, and we have shown that for any j and t, $F_j(t) \xrightarrow{p} \Phi_{\sigma^2/(1-\kappa)}(t)$ in the proof of Theorem 1.

   (b) This is true because β_j's are i.i.d. We will find $\mathcal{L}(T^{(1)})$ in part (c).

   (c) Since

$$\begin{pmatrix} \sqrt{n}\hat{\beta}_j \\ \sqrt{n}\hat{\beta}_k \end{pmatrix} - \begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{pmatrix} \mid \mathbf{X} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \sigma^2 \left[\left(\frac{\mathbf{X}^\top \mathbf{X}}{n} \right)^{-1} \right]_{\{j,k\},\{j,k\}} \right),$$

   we can show

$$\begin{pmatrix} \sqrt{n}\hat{\beta}_j \\ \sqrt{n}\hat{\beta}_k \end{pmatrix} - \begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\sigma^2}{1-\kappa} I_2 \right).$$

   once we verify that any 2×2 sub-diagonal matrix of $\left(\mathbf{X}^\top\mathbf{X}/n\right)^{-1}$ converges in probability to $(1-\kappa)^{-1}I_2$, which follows directly from a computation of the first and second moments of the inverse Wishart distribution.

   To sum up, the p-values we obtain from OLS satisfy the conditions of Theorem 10 with $F^{(1)} = \Phi_{\sigma^2/(1-\kappa)}$ and $T^{(1)} \sim B_0 + \sigma W/\sqrt{1-\kappa}$, where W is a standard Gaussian random variable independent of B_0. We obtain the effective π_μ by dividing the T_j's by $\sigma/\sqrt{1-\kappa}$.

3. *Distilled lasso.* Let $T_j = (\mathbf{Y} - \mathbf{X}_{-j}\hat{\beta}_\lambda^{(-j)})^\top \mathbf{X}_j$, where $\hat{\beta}_\lambda^{(-j)}$ is the lasso coefficient of fitting lasso with parameter λ on $\mathbf{Y}$ against $\mathbf{X}_{-j}$. Our F_j in this case is the CDF of $\mathcal{N}(0, \|\mathbf{Y} - \mathbf{X}_{-j}\hat{\beta}_\lambda^{(-j)}\|_2^2/n)$. We now check the conditions for Theorem 10.

   (a) It is clear that $1 - F_j(T_j) \sim \text{Unif}[0,1]$ for the null variables. By Lemma 6, we have for any j, $F_j(t) \xrightarrow{p} \Phi_{\lambda^2/\alpha_\lambda^2}(t)$. We will find $\mathcal{L}(T^{(1)})$ in part (c).

   (b) This is true because β_j's are i.i.d.

   (c) We verify this condition by introducing Lemma 10.

   Lemma 10. *In Setting 2, for distinct j and k we have*

$$\begin{pmatrix} T_j \\ T_k \end{pmatrix} - \frac{\lambda}{\alpha_\lambda \tau_\lambda} \begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\lambda^2}{\alpha_\lambda^2} I_2 \right).$$

   To sum up, the p-values we obtain from the distilled lasso statistic satisfy the conditions of Theorem 10 with $F^{(1)} = \Phi_{\lambda^2/\alpha_\lambda^2}$ and $T^{(1)} \sim \lambda B_0/(\alpha_\lambda\tau_\lambda) + \lambda W/\alpha_\lambda$, where W is a standard Gaussian random variable independent of B_0. We obtain the effective π_μ by dividing the T_j's by λ/α_λ.

□

Lemma 11. *In Setting 2, let $(X^\top, Y)$ represent a random row vector that has the same distribution as one generic row of $[\mathbf{X}, \mathbf{Y}]$ conditional on β, then*

$$X \mid Y, \beta \sim \mathcal{N}\left(\frac{Y}{\|\beta\|^2+\sigma^2}\beta, I - \frac{1}{\|\beta\|^2+\sigma^2}\beta\beta^\top\right) \tag{21}$$

and

$$\mathbf{X}^\top\mathbf{Y} \mid \mathbf{Y}, \beta \sim \mathcal{N}\left(\frac{\|\mathbf{Y}\|^2}{\|\beta\|^2+\sigma^2}\beta, \|\mathbf{Y}\|^2\left(I - \frac{1}{\|\beta\|^2+\sigma^2}\beta\beta^\top\right)\right). \tag{22}$$

Proof of Lemma 11. Jointly, we have

$$\begin{pmatrix} X \\ Y \end{pmatrix} \mid \beta \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} I_p & \beta \\ \beta^\top & \|\beta\|_2^2+\sigma^2 \end{bmatrix}\right).$$

Apply the formula of the conditional Gaussian distribution and we get Equation (21). Thus,

$$YX \mid Y, \beta \sim \mathcal{N}\left(\frac{Y^2}{\|\beta\|^2+\sigma^2}\beta, Y^2\left(I - \frac{1}{\|\beta\|^2+\sigma^2}\beta\beta^\top\right)\right).$$

We notice that the left hand side of Equation (22) is just a summation of n independent Gaussian random vectors with their distributions given by the above formula, and the validity of Equation (22) then follows. □

Proof of Lemma 9. Let $T_j = n^{-1/2}\mathbf{X}_j^\top\mathbf{Y}$. Applying Lemma 11, we have that for $j \neq k$,

$$\begin{pmatrix} T_j - \sqrt{n}\beta_j \\ T_k - \sqrt{n}\beta_k \end{pmatrix} \mid \mathbf{Y}, \beta$$
$$\sim \mathcal{N}\left(\left(\frac{\|\mathbf{Y}\|^2/n}{\|\beta\|^2+\sigma^2} - 1\right)\begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{pmatrix}, \frac{\|\mathbf{Y}\|^2}{n}\begin{bmatrix} 1 - \frac{\beta_j^2}{\|\beta\|^2+\sigma^2} & -\frac{\beta_j\beta_k}{\|\beta\|^2+\sigma^2} \\ -\frac{\beta_j\beta_k}{\|\beta\|^2+\sigma^2} & 1 - \frac{\beta_k^2}{\|\beta\|^2+\sigma^2} \end{bmatrix}\right),$$

which converges in distribution to $\mathcal{N}(0, (\sigma^2 + \kappa\mathbb{E}[B_0^2])I_2)$ because

$$\|\mathbf{Y}\|^2/n \xrightarrow{p} \sigma^2 + \kappa\mathbb{E}[B_0^2],$$
$$\|\beta\|^2 \xrightarrow{p} \kappa\mathbb{E}[B_0^2],$$
$$\beta_j^2, \beta_k^2, \beta_j\beta_k \xrightarrow{p} 0,$$

and $\sqrt{n}\beta_j$'s are universally bounded.

□

Proof of Lemma 10. To use the results in Bayati and Montanari (2011), we apply the following re-normalization to Setting 2: assume $\mathbf{X}$ is divided by $\sqrt{n}$ and β is multiplied by $\sqrt{n}$. As explained in the proof of Lemma 6, we additionally assume that ε_i's and X_{ij}'s do not change with n, p as long as $n \geq i$ and $p \geq j$, which does not change the distribution of Setting 2 for each fixed pair (n, p).

Let $\hat{\mathbf{Y}}^{(\text{-}j)} = \mathbf{X}_{\text{-}j}\hat{\beta}_\lambda^{(\text{-}j)}$, where $\mathbf{X}_{\text{-}j}$ is $\mathbf{X}$ with the jth column removed, and $\hat{\beta}_\lambda^{(\text{-}j)}$ is the lasso coefficient from regressing $\mathbf{Y}$ on $\mathbf{X}_{\text{-}j}$ with penalty parameter λ. We replace j, k with $1, 2$ in the proof, which we can do due to exchangeability. Note that

$$\begin{pmatrix} T_1 \\ T_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}1)})^\top\mathbf{X}_1 \\ (\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}2)})^\top\mathbf{X}_2 \end{pmatrix}.$$

We first consider the statistic

$$\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix} = \begin{pmatrix} (\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}(1:2))})^\top \mathbf{X}_1 \\ (\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}(1:2))})^\top \mathbf{X}_2 \end{pmatrix},$$

where $\hat{\mathbf{Y}}^{(\text{-}(1:2))} = \mathbf{X}_{\text{-}(1:2)} \hat{\beta}_\lambda^{(\text{-}(1:2))}$, $\mathbf{X}_{\text{-}(1:2)}$ is $\mathbf{X}$ with its first two columns removed, and $\hat{\beta}_\lambda^{(\text{-}(1:2))}$ is the lasso coefficient from regressing $\mathbf{Y}$ on $\mathbf{X}_{\text{-}(1:2)}$ with penalty parameter λ.

Consider a random row vector $(X_1, X_2, X_{\text{-}(1:2)}^\top, Y)$ that has the same distribution of a generic row of $[\mathbf{X}, \mathbf{Y}]$. Applying Lemma 12 which we will introduce shortly, we have (note the re-normalization at the beginning of this proof)

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \mid X_{\text{-}(1:2)}, Y, \beta \sim \mathcal{N}\left(\frac{Y - X_{\text{-}(1:2)}^\top \beta_{\text{-}(1:2)}}{n\sigma^2 + \beta_1^2 + \beta_2^2} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \frac{1}{n} \begin{bmatrix} 1 - \frac{\beta_1^2}{n\sigma^2+\beta_1^2+\beta_2^2} & -\frac{\beta_1\beta_2}{n\sigma^2+\beta_1^2+\beta_2^2} \\ -\frac{\beta_1\beta_2}{n\sigma^2+\beta_1^2+\beta_2^2} & 1 - \frac{\beta_2^2}{n\sigma^2+\beta_1^2+\beta_2^2} \end{bmatrix} \right).$$

It is then easy to see that (by writing $(\tilde{T}_1, \tilde{T}_2)$ as a sum of n independent Gaussian random vectors)

$$\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix} \mid \mathbf{Y}, \mathbf{X}_{\text{-}(1:2)}, \beta \sim$$
$$\mathcal{N}\left(\frac{(\mathbf{Y} - \mathbf{X}_{\text{-}(1:2)} \hat{\beta}_\lambda^{(\text{-}(1:2))})^\top \varepsilon'}{n\sigma^2 + \beta_1^2 + \beta_2^2} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \frac{\|\mathbf{Y} - \mathbf{X}_{\text{-}(1:2)} \hat{\beta}_\lambda^{(\text{-}(1:2))}\|^2}{n} \begin{bmatrix} 1 - \frac{\beta_1^2}{n\sigma^2+\beta_1^2+\beta_2^2} & -\frac{\beta_1\beta_2}{n\sigma^2+\beta_1^2+\beta_2^2} \\ -\frac{\beta_1\beta_2}{n\sigma^2+\beta_1^2+\beta_2^2} & 1 - \frac{\beta_2^2}{n\sigma^2+\beta_1^2+\beta_2^2} \end{bmatrix} \right),$$

where $\varepsilon' = \mathbf{Y} - \mathbf{X}_{\text{-}(1:2)} \beta_{\text{-}(1:2)} = \varepsilon + \beta_1 \mathbf{X}_1 + \beta_2 \mathbf{X}_2$ is the effective error in the model $\mathbf{Y} \sim \mathbf{X}_{\text{-}(1:2)}$. Using the results from the proof of Lemma 6, we can find the limits of

$$\frac{(\mathbf{Y} - \mathbf{X}_{\text{-}(1:2)} \hat{\beta}_\lambda^{(\text{-}(1:2))})^\top \varepsilon'}{n} \quad \text{and} \quad \frac{\|\mathbf{Y} - \mathbf{X}_{\text{-}(1:2)} \hat{\beta}_\lambda^{(\text{-}(1:2))}\|^2}{n},$$

and see that

$$\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix} - \frac{\lambda}{\alpha_\lambda \tau_\lambda} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \frac{\lambda^2}{\alpha_\lambda^2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right),$$

which is a bivariate Gaussian distribution with i.i.d. components. Now we just have to show

$$\begin{pmatrix} \tilde{T}_1 \\ \tilde{T}_2 \end{pmatrix} - \begin{pmatrix} T_1 \\ T_2 \end{pmatrix} \xrightarrow{d} 0.$$

It suffices to show $T_1 - \tilde{T}_1 \xrightarrow{d} 0$ marginally (same for $T_2 - \tilde{T}_2$ because of symmetry). Note that $T_1 - \tilde{T}_1 = \left(\hat{\mathbf{Y}}^{(\text{-}(1:2))} - \hat{\mathbf{Y}}^{(-1)}\right)^\top \mathbf{X}_1$, and we can again apply Lemma 12 to get $\mathcal{L}(X_1 \,|\, X_{\text{-}1}, Y, \beta)$, where $(X_1, X_{\text{-}1}^\top, Y)$ has the same distribution of a generic row of $[\mathbf{X}, \mathbf{Y}]$. We would get

$$T_1 - \tilde{T}_1 \mid \mathbf{Y}, \mathbf{X}_{(\text{-}1)}, \beta \sim \mathcal{N}\left(\frac{\beta_1}{n\sigma^2 + \beta_1^2} (\hat{\mathbf{Y}}^{(\text{-}1)} - \hat{\mathbf{Y}}^{(\text{-}1:2)})^\top (\varepsilon + \beta_1 \mathbf{X}_1), \frac{\sigma^2}{n\sigma^2 + \beta_1^2} \|\hat{\mathbf{Y}}^{\text{-}1} - \hat{\mathbf{Y}}^{(\text{-}1:2)}\|^2 \right).$$

Now it remains to show $\|\hat{\mathbf{Y}}^{(\text{-}1)} - \hat{\mathbf{Y}}^{(\text{-}1:2)}\|^2 / n \xrightarrow{p} 0$, which would imply the above variance and mean (use Cauchy–Schwartz) both go to zero. Note that we can simplify this problem to $\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{(\text{-}1)}\|^2 / n \xrightarrow{p} 0$, because both regression processes ignore the first column of $\mathbf{X}$ and we can just treat $\beta_1 \mathbf{X}_1$ as part of the error vector, which does not change the asymptotic distribution of the error.

Line (d) of the first decomposition used in the proof of Lemma 3.1 in Bayati and Montanari (2011) (we take their x to be our $\hat{\beta}_\lambda^{(\text{-}1)}$ and their r to be our $\hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)}$. Here, we slightly abuse notation: $\hat{\beta}_\lambda^{(\text{-}1)}$ is originally $(p-1)$-dimensional, and we add a zero as its first coordinate to make it comparable with $\hat{\beta}_\lambda$) shows that the sum of four terms is non-positive, and immediately after it is shown that three of those terms, including $\frac{\|\hat{\mathbf{Y}}-\hat{\mathbf{Y}}^{(\text{-}1)}\|^2}{2p}$ (their A is our $\mathbf{X}$, and $\mathbf{X}\hat{\beta}_\lambda = \hat{\mathbf{Y}}$ and $\mathbf{X}\hat{\beta}_\lambda^{(\text{-}1)} = \hat{\mathbf{Y}}^{(\text{-}1)}$), are non-negative, guaranteeing that the remaining term, $\langle \text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)}), \hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)} \rangle$ is negative and has absolute value greater than each of the three non-negative terms. Thus:

$$\frac{\|\hat{\mathbf{Y}} - \hat{\mathbf{Y}}^{(\text{-}1)}\|^2}{2p} \leq |\langle \text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)}), \hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)} \rangle|, \tag{23}$$

where $\text{sg}(\mathcal{C}, \beta)$ is any subgradient of $\mathcal{C}(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2/2 + \lambda\|\beta\|_1$, i.e.,

$$-\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta) + \lambda\gamma \tag{24}$$

for a γ that is a subgradient of the p-dimensional L_1 norm. By the Cauchy–Schwarz inequality,

$$|\langle \text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)}), \hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)} \rangle| \leq \sqrt{\frac{\|\text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)})\|^2}{p}} \sqrt{\frac{\|\hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)}\|^2}{p}}. \tag{25}$$

Since $\hat{\beta}_\lambda^{(\text{-}1)}$ is a lasso solution, the Karush–Kuhn–Tucker (KKT) conditions imply

$$\mathbf{X}_{\text{-}1}^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}_\lambda^{(-1)}) = \lambda\gamma^*, \tag{26}$$

where for $j = 1, \ldots, p-1$, (note again that we have added a zero to $\hat{\beta}_\lambda^{(\text{-}1)}$ to make it p-dimensional)

$$\gamma_j^* \in \begin{cases} \{1\}, & (j+1)\text{st coordinate of } \hat{\beta}_\lambda^{(\text{-}1)} > 0, \\ \{-1\}, & (j+1)\text{st coordinate of } \hat{\beta}_\lambda^{(\text{-}1)} < 0, \\ [-1, 1], & (j+1)\text{st coordinate of } \hat{\beta}_\lambda^{(\text{-}1)} = 0. \end{cases} \tag{27}$$

Since the first coordinate of $\hat{\beta}_\lambda^{(\text{-}1)}$ is zero, directly from the definition of a subgradient (24), the first coordinate of $\text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)})$ can be any number in

$$\left[-\mathbf{X}_1^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}_\lambda^{(\text{-}1)}) - \lambda, -\mathbf{X}_1^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}_\lambda^{(\text{-}1)}) + \lambda\right]. \tag{28}$$

For the remaining $(p-1)$ coordinates, they can be $-\mathbf{X}_{\text{-}1}^\top(\mathbf{Y} - \mathbf{X}\hat{\beta}_\lambda^{(\text{-}1)}) + \lambda\gamma$ for any γ that satisfies (27) (i.e., a subgradient of the $(p-1)$-dimensional L_1 norm at $\hat{\beta}_\lambda^{(\text{-}1)}$), and specifically, we can let γ be the one that satisfies (26), so that the subgradient in these $(p-1)$ dimensions cancels to 0. This way, we have defined a $\text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)})$ so that its first coordinate is (take the midpoint of (28))

$$-(\mathbf{Y} - \underbrace{\hat{\mathbf{Y}}^{(\text{-}1)}}_{=\mathbf{X}\hat{\beta}_\lambda^{(\text{-}1)}})^\top \mathbf{X}_1 \tag{29}$$

and all other coordinates are zero. Note that

$$\mathcal{L}\left(-(\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}1)})^\top \mathbf{X}_1 \mid \mathbf{Y}, \mathbf{X}_{\text{-}1}\right) = \mathcal{N}(0, \|\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}1)}\|^2/n)$$

and $\|\mathbf{Y} - \hat{\mathbf{Y}}^{(\text{-}1)}\|^2/n \overset{\text{a.s.}}{\to} \lambda^2/\alpha_\lambda^2$ by Lemma 6. Thus, (29) converges to $\mathcal{N}(0, \lambda^2/\alpha_\lambda^2)$ in distribution. This way, the squared L_2 norm of the selected $\text{sg}(\mathcal{C}, \hat{\beta}_\lambda^{(\text{-}1)})$ divided by p converges to zero. On the other hand,

$$\frac{\|\hat{\beta}_\lambda - \hat{\beta}_\lambda^{(\text{-}1)}\|^2}{p} \le \frac{2}{p}(\|\hat{\beta}_\lambda\|^2 + \|\hat{\beta}_\lambda^{(\text{-}1)}\|^2),$$

and the right hand side converges to a constant as a corollary of Theorem 1.5 in Bayati and Montanari (2011). Now we get $\|\hat{\mathbf{Y}}^{(\text{-}1)} - \hat{\mathbf{Y}}^{(\text{-}1:2)}\|^2/n \overset{p}{\to} 0$ from (23) and (25), because n/p converges to a positive constant. □

Lemma 12. *Let W_1 and W_2 be independent q-dimensional and r-dimensional standard multivariate Gaussian random vectors. Let $Y \mid W_1, W_2 \sim \mathcal{N}(W_1^\top \zeta_q + W_2^\top \zeta_r, \sigma^2)$. Then,*

$$W_1 \mid \begin{pmatrix} W_2 \\ Y \end{pmatrix} \sim \mathcal{N}\left(\frac{Y - W_2^\top \zeta_r}{\sigma^2 + \|\zeta_q\|_2^2}\zeta_q, I_q - \frac{1}{\sigma^2 + \|\zeta_q\|_2^2}\zeta_q^\top \zeta_q\right). \tag{30}$$

Proof of Lemma 12. Jointly, we have

$$\begin{pmatrix} W_1 \\ W_2 \\ Y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} I_q & 0 & \zeta_q \\ 0 & I_r & \zeta_r \\ \zeta_q^\top & \zeta_r^\top & \|\zeta_q\|_2^2 + \|\zeta_r\|_2^2 + \sigma^2 \end{bmatrix}\right).$$

Note that

$$\begin{bmatrix} I_r & \zeta_r \\ \zeta_r^\top & \|\zeta_q\|_2^2 + \|\zeta_r\|_2^2 + \sigma^2 \end{bmatrix}^{-1} = \begin{bmatrix} I_r + \frac{\zeta_r \zeta_r^\top}{\sigma^2 + \|\zeta_q\|_2^2} & -\frac{\zeta_r}{\sigma^2 + \|\zeta_q\|_2^2} \\ -\frac{\zeta_r^\top}{\sigma^2 + \|\zeta_q\|_2^2} & \frac{1}{\sigma^2 + \|\zeta_q\|_2^2} \end{bmatrix}.$$

Thus, directly apply the formula for the conditional Gaussian distribution and we have (30). □

Theorem 11. *Let J_0 and J_1 form a partition of $\{1, 2, \ldots, p\}$ and $|J_0|/p \to \gamma \in (0, 1)$. Consider p random variables W_j, which can be thought of as the W_j's in a knockoffs procedure. Assume the following conditions.*

1. *For $j \in J_0$, $W_j \overset{d}{\to} W^{(0)} \sim G^{(0)}$ and for $j \in J_1$, $W_j \overset{d}{\to} W^{(1)} \sim G^{(1)}$. $G^{(0)}$ and $G^{(1)}$ are deterministic CDFs of random variables with a common support which is connected and symmetric around 0, and continuous densities on that support.*
2. *Within J_0 or J_1, the W_j's are exchangeable.*
3. *For distinct $j_1, j_2 \in J_0$ and distinct $j_3, j_4 \in J_1$, the following two pairs of random variables are asymptotically pairwise independent: (W_{j_1}, W_{j_2}) and (W_{j_3}, W_{j_4}). That is, both pairs converge in distribution to a bivariate random vector (not necessarily the same random vector) with independent components.*

Let (min over an empty set is defined to be infinity)

$$w_{\text{KF}} = \min\{w \ge 0 : g(w) \le q\},$$

where

$$g(w) = \frac{\gamma G^{(0)}(-w) + (1-\gamma)(G^{(1)}(-w))}{\gamma(1 - G^{(0)}(w)) + (1-\gamma)(1 - G^{(1)}(w))}.$$

Then for almost every $q \in (0,1)$, at least one of the following cases is true: (i) $w_{\mathrm{KF}} > 0$, $g'(w_{\mathrm{KF}}) \neq 0$, (ii) $w_{\mathrm{KF}} = 0$, $g'(0) < 0$, (iii) $w_{\mathrm{KF}} = 0$, $g(0) < q$, or (iv) $w_{\mathrm{KF}} = \infty$. In cases (i), (ii), or (iii), for the knockoff filter (2) at level q applied to $W_1, \ldots, W_p$, the FDP and realized power converge in probability to

$$\frac{\gamma G^{(0)}(-w_{\mathrm{KF}})}{\gamma(1-G^{(0)}(w_{\mathrm{KF}})) + (1-\gamma)(1-G^{(1)}(w_{\mathrm{KF}}))} \quad \textit{and} \quad 1 - G^{(1)}(w_{\mathrm{KF}}),$$

respectively. In case (iv), the realized power converges in probability to 0.

Proof of Theorem 11. If $q > g(0)$, then (iii) holds. If $q < \inf\{g(w) : w \geq 0\}$, then (iv) holds. If $q \in (\inf\{g(w) : w \geq 0\}, g(0))$, then because g is continuous, we can see that $w_{\mathrm{KF}} \in (0, \infty)$ (note that we are considering minimum and $[0, \infty)$ is closed on the left, so the minimum must exist and not equal to 0). And in this case, $g(w_{\mathrm{KF}}) = q$, since otherwise $g(w_{\mathrm{KF}}) < q$ and w_{KF} could be smaller because of g's continuity. Next we show for almost every $q \in (\inf\{g(w) : w \geq 0\}, g(0))$, condition (i) is met. We just need to show that the set $\{g(w) : g'(w) = 0\}$ has measure zero, which is a simple application of Sard's theorem (Lemma 7, take $n = m = k = 1$, $f(x) = g(a \arctan(x))$ for a suitable a such that $a \arctan(x)$ matches the support of $G^{(0)}$ and $G^{(1)}$ when that support is finite; otherwise just take $f = g$), where g' is continuous because $G^{(0)}$ and $G^{(1)}$ have continuous densities with a common support.

In case (i), we have established that $g(w_{\mathrm{KF}}) = q$. Since $g'(w_{\mathrm{KF}}) \neq 0$, we must have $g'(w_{\mathrm{KF}}) < 0$, since otherwise w_{KF} could also be smaller. Thus, in cases (i) and (ii), $g'(w_{\mathrm{KF}}) < 0$ and for any sufficiently small $\varepsilon > 0$ there exists a point w^* in $(w_{\mathrm{KF}}, w_{\mathrm{KF}} + \varepsilon)$ such that $g(w^*) < q$. In case (iii), since g is continuous and $g(0) < q$, it also holds that for any sufficiently small $\varepsilon > 0$ there exists a point w^* in $(w_{\mathrm{KF}}, w_{\mathrm{KF}} + \varepsilon)$ such that $g(w^*) < q$.

Let

$$\hat{w}_p = \min\{w > 0 : g_p(w) \equiv \frac{1/p + \#\{j : W_j \leq -w\}/p}{\#\{j : W_j \geq w\}/p} \leq q\}.^8$$

We analyze cases (i), (ii), and (iii). We begin by showing $\hat{w}_p \xrightarrow{p} w_{\mathrm{KF}}$. Take any sufficiently small $\varepsilon > 0$. We have established that there exists a point w^* in $(w_{\mathrm{KF}}, w_{\mathrm{KF}} + \varepsilon)$ such that $g(w^*) < q$. Then

$$\begin{aligned}
\mathbb{P}(\hat{w}_p < w_{\mathrm{KF}} + \varepsilon) &\geq \mathbb{P}(\hat{w}_p \leq w^*) \\
&\geq \mathbb{P}(g_p(w^*) \leq q) \\
&\geq \mathbb{P}(|g_p(w^*) - g(w^*)| < |g(w^*) - q|) \to 1.
\end{aligned}$$

In case (iii), we get $\mathbb{P}(\hat{w}_p \geq w_{\mathrm{KF}} = 0) = 1$ for free and the proof is concluded. Next we consider cases (i) and (ii) and assume $\varepsilon < w_{\mathrm{KF}}$. Choose $\delta_1 \in \big(0, \min(1 - G^{(0)}(w_{\mathrm{KF}} - \varepsilon), 1 - G^{(1)}(w_{\mathrm{KF}} - \varepsilon))\big)$. Let $\delta_3 = \min\{g(w) - q : 0 \leq w \leq w_{\mathrm{KF}} - \varepsilon\}$, which is positive since otherwise g could attain a value no more than q in $[0, w_{\mathrm{KF}} - \varepsilon]$, violating w_{KF}'s definition. Now observe that the function $\frac{\gamma x + (1-\gamma)y}{\gamma z + (1-\gamma)t}$ is continuous in (x, y, z, t) on $\{(x, y, z, t) \in [0,1]^4 : z, t \geq \min(1 - G^{(0)}(w_{\mathrm{KF}} - \varepsilon), 1 - G^{(1)}(w_{\mathrm{KF}} - \varepsilon)) - \delta_1\}$, and thus uniformly continuous. Choose $\delta_2 \in (0, \delta_1)$ such that whenever $(x, y, z, t), (x', y', z', t') \in \{(x, y, z, t) \in [0,1]^4 : z, t \geq \min(1 - G^{(0)}(\tau_{\mathrm{KF}} - \varepsilon), 1 - G^{(1)}(\tau_{\mathrm{KF}} - \varepsilon))\}$ and $|x - x'|, |y - y'|, |z - z'|, |t - t'| < \delta_2$, $|\frac{\gamma x + (1-\gamma)y}{\gamma z + (1-\gamma)t} - \frac{\gamma x' + (1-\gamma)y'}{\gamma z' + (1-\gamma)t'}| < \delta_3$.

By Lemma 8, with probability converging to 1, all the CDFs fall within a δ_2-neighborhood around the limit CDFs, and by the previously shown uniform continuity, $|g_p(w) - g(w)| < \delta_3$ for

[8] Formally, we only take the minimum over $w \in \{|W_j| : W_j \neq 0\}$. Such a difference is important when $g_p(w) \leq q$ for all $w > 0$. The term $1/p$ does not matter asymptotically, in that we can still consider the numerator as an empirical CDF of the W_j's.

$w \in [0, w_{\mathrm{KF}} - \varepsilon]$. By the definition of δ_3, this means with probability converging to 1, $g_p(w) > q$ for all $w \in [0, w_{\mathrm{KF}} - \varepsilon]$. Now we have

$$\mathbb{P}(\hat{w}_p > w_{\mathrm{KF}} - \varepsilon) = \mathbb{P}(g_p(w) > q, \forall w \in [0, w_{\mathrm{KF}} - \varepsilon]) \to 1.$$

Combining the results, we have

$$\lim_{p\to\infty} \mathbb{P}(|\hat{w}_p - w_{\mathrm{KF}}| < \varepsilon) = 1.$$

Similar to the proof of Theorem 10, we can show that

$$\frac{\#\{j \in J_0 : W_j \le t\}}{|J_0|} \xrightarrow{p} G^{(0)}(t) \quad \text{and} \quad \frac{\#\{j \in J_1 : W_j \le t\}}{|J_1|} \xrightarrow{p} G^{(1)}(t),$$

and the convergence is uniform over $t \in \mathbb{R}$ by Lemma 8. The result then follows by noticing that the FDP and realized power are

$$\frac{\#\{j \in J_0 : W_j \ge \hat{w}_p\}}{\#\{j : W_j \ge \hat{w}_p\}} \quad \text{and} \quad \frac{\#\{j \in J_1 : W_j \ge \hat{w}_p\}}{|J_1|},$$

respectively.

Last, we look at case (iv). Similar to the end of the proof of Theorem 10, we can show that in this case the asymptotic realized power is 0 by showing that for any $M > 0$, $\mathbb{P}(w_{\mathrm{KF}} \ge M) \to 1$. □

Lemma 13. *Let J_0 and J_1 form a partition of $\{1, 2, \ldots, p\}$ and $|J_0|/p \to \gamma \in (0, 1)$. If $W_j = f(T_j, \tilde{T}_j)$ for a continuous antisymmetric function f, then the following conditions imply conditions 1, 2, and 3 of Theorem 11.*

1. *For $j \in J_0$, $(T_j, \tilde{T}_j) \xrightarrow{d} T^{(0)}$ and for $j \in J_1$, $(T_j, \tilde{T}_j) \xrightarrow{d} T^{(1)}$. $T^{(0)}$, $T^{(1)}$, and f are such that the distributions of $f(T^{(0)})$ and $f(T^{(1)})$ have a common support and continuous densities.*
2. *Within J_0 or J_1, the $(T_j, \tilde{T}_j)$'s are exchangeable.*
3. *For distinct $j_1, j_2 \in J_0$ and distinct $j_3, j_4 \in J_1$, the following two pairs of random vectors are asymptotically pairwise independent:*

$$\left(\begin{pmatrix} T_{j_1} \\ \tilde{T}_{j_1} \end{pmatrix}, \begin{pmatrix} T_{j_2} \\ \tilde{T}_{j_2} \end{pmatrix}\right) \quad \text{and} \quad \left(\begin{pmatrix} T_{j_3} \\ \tilde{T}_{j_3} \end{pmatrix}, \begin{pmatrix} T_{j_4} \\ \tilde{T}_{j_4} \end{pmatrix}\right).$$

The proof of Lemma 13 is immediate and thus omitted.

Theorem 6. *In Setting 2, for almost every $q \in (0, 1)$, knockoffs with $\tilde{\mathbf{X}}$ an i.i.d. copy of $\mathbf{X}$ and the antisymmetric function $f(x, y) = x - y$ at level q with marginal covariance or OLS test statistic has the following one-sided effective π_μ's with respect to the AdaPT procedure at level q:*

1. *For the marginal covariance statistic, the effective π_μ is the distribution of $\frac{1}{\sqrt{2(\sigma^2 + \kappa\mathbb{E}[B_0^2])}} B_0$.*
2. *For the OLS statistic, assuming $\kappa < 1/2$, the effective π_μ is the distribution of $\frac{\sqrt{1-2\kappa}}{\sqrt{2\sigma^2}} B_0$.*

Proof of Theorem 6. Similar to the proof of Theorem 5, we analyze the two statistics separately.

1. *Marginal covariance.* Consider using $T_j = n^{-1/2}\mathbf{X}_j^\top \mathbf{Y}$ and $\tilde{T}_j = n^{-1/2}\tilde{\mathbf{X}}_j^\top \mathbf{Y}$. To utilize Lemma 13, we introduce Lemma 14. The proof is based on a tedious yet straightforward computation of the characteristic function of the Wishart distribution.

Lemma 14. *In Setting 2 with the knockoffs procedure that takes* $\tilde{\mathbf{X}}$ *to be an i.i.d. copy of* $\mathbf{X}$*, let* $T_j = n^{-1/2}\mathbf{X}_j^\top \mathbf{Y}$ *and* $\tilde{T}_j = n^{-1/2}\tilde{\mathbf{X}}_j^\top \mathbf{Y}$*. We have for* $j \neq k$*,*

$$\begin{pmatrix} T_j - \sqrt{n}\beta_j \\ T_k - \sqrt{n}\beta_k \\ \tilde{T}_j \\ \tilde{T}_k \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(0, \left(\sigma^2 + \kappa\mathbb{E}[B_0^2]\right) I_4\right).$$

This means that asymptotically, we can think of all the T_j's and $\tilde{T}_j$'s as independent, $\tilde{T}_j \sim \mathcal{N}(0, \sigma^2 + \kappa\mathbb{E}[B_0^2])$ and $T_j \sim B_0 + \sqrt{\sigma^2 + \kappa\mathbb{E}[B_0^2]}Z$, where $Z \sim \mathcal{N}(0,1)$ is independent of B_0. Thus, we can think of $W_j = T_j - \tilde{T}_j$ and $W_k = T_k - \tilde{T}_k$ as independent for distinct j and k, with distribution $B_0 + \sqrt{2(\sigma^2 + \kappa\mathbb{E}[B_0^2])}Z$, where $Z \sim \mathcal{N}(0,1)$ is independent of B_0. We obtain the effective π_μ by dividing the W_j's by $\sqrt{2(\sigma^2 + \kappa\mathbb{E}[B_0^2])}$.

2. *OLS.* We consider letting $T_j = \sqrt{n}\hat{\beta}_j$ and $\tilde{T}_j = \sqrt{n}\hat{\beta}_{j+p}$, where $\kappa < 1/2$ and $\hat{\beta}$ is the OLS coefficient of $\mathbf{Y}$ against $[\mathbf{X}, \tilde{\mathbf{X}}]$. We just check the conditions in Lemma 13. The second condition is obvious. For the other two conditions, notice that for $j \neq k$,

$$\begin{pmatrix} \sqrt{n}\hat{\beta}_j \\ \sqrt{n}\hat{\beta}_{j+p} \\ \sqrt{n}\hat{\beta}_k \\ \sqrt{n}\hat{\beta}_{k+p} \end{pmatrix} - \begin{pmatrix} \sqrt{n}\beta_j \\ 0 \\ \sqrt{n}\beta_k \\ 0 \end{pmatrix} \mid \mathbf{X}, \tilde{\mathbf{X}} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \sigma^2 \left[\begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top\mathbf{X} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \end{pmatrix}^{-1}\right]_{(j,j+p,k,k+p),(j,j+p,k,k+p)}\right).$$

Thus, we can show

$$\begin{pmatrix} \sqrt{n}\hat{\beta}_j \\ \sqrt{n}\hat{\beta}_{j+p} \\ \sqrt{n}\hat{\beta}_k \\ \sqrt{n}\hat{\beta}_{k+p} \end{pmatrix} - \begin{pmatrix} \sqrt{n}\beta_j \\ 0 \\ \sqrt{n}\beta_k \\ 0 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \frac{\sigma^2}{1-2\kappa} I_4\right),$$

once we verify that any 4×4 sub-diagonal matrix of

$$\begin{pmatrix} \mathbf{X}^\top\mathbf{X} & \mathbf{X}^\top\tilde{\mathbf{X}} \\ \tilde{\mathbf{X}}^\top\mathbf{X} & \tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} \end{pmatrix}^{-1}$$

converges in probability to $(1-2\kappa)^{-1}I_4$, which follows directly from a computation of the first and second moments of the inverse Wishart distribution. This means that asymptotically, we can think of all the T_j's and $\tilde{T}_j$'s as independent, $\tilde{T}_j \sim \mathcal{N}(0, \sigma^2/(1-2\kappa))$ and $T_j \sim B_0 + \sigma Z/\sqrt{1-2\kappa}$, where $Z \sim \mathcal{N}(0,1)$ is independent of B_0. Thus, we can think of $W_j = T_j - \tilde{T}_j$ and $W_k = T_k - \tilde{T}_k$ as independent for distinct j and k, with distribution $B_0 + \sqrt{2}\sigma Z/\sqrt{1-2\kappa}$, where $Z \sim \mathcal{N}(0,1)$ is independent of B_0. We obtain the effective π_μ by dividing the W_j's by $\sqrt{2}\sigma/\sqrt{1-2\kappa}$.

□

Proof of Lemma 14. By Lemma 11,

$$\mathbf{X}^\top\mathbf{Y} \mid \mathbf{Y}, \beta \sim \mathcal{N}\left(\frac{\|\mathbf{Y}\|^2}{\|\beta\|^2 + \sigma^2}\beta, \|\mathbf{Y}\|^2\left(I - \frac{1}{\|\beta\|^2 + \sigma^2}\beta\beta^\top\right)\right),$$

and since $\tilde{\mathbf{X}}^\top \mathbf{Y} \mid \mathbf{Y}, \beta \sim \mathcal{N}(0, \|\mathbf{Y}\|^2 I_p)$, $\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{X} \mid \mathbf{Y}$,

$$\begin{pmatrix} \mathbf{X}^\top \mathbf{Y} \\ \tilde{\mathbf{X}}^\top \mathbf{Y} \end{pmatrix} \mid \mathbf{Y}, \beta \sim \mathcal{N}\left(\frac{\|\mathbf{Y}\|^2}{\|\beta\|^2 + \sigma^2} \begin{pmatrix} \beta \\ 0 \end{pmatrix}, \|\mathbf{Y}\|^2 \begin{bmatrix} I - \frac{1}{\|\beta\|^2+\sigma^2} \beta\beta^\top & 0 \\ 0 & I \end{bmatrix} \right)$$

Let $T_j = n^{-1/2} \mathbf{X}_j^\top \mathbf{Y}$ and $\tilde{T}_j = n^{-1/2} \tilde{\mathbf{X}}_j^\top \mathbf{Y}$. For $j \neq k$,

$$\begin{pmatrix} T_j - \sqrt{n}\beta_j \\ T_k - \sqrt{n}\beta_k \\ \tilde{T}_j \\ \tilde{T}_k \end{pmatrix} \mid \mathbf{Y}, \beta$$

$$\sim \mathcal{N}\left(\left(\frac{\|\mathbf{Y}\|^2/n}{\|\beta\|^2 + \sigma^2} - 1 \right) \begin{pmatrix} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \\ 0 \\ 0 \end{pmatrix}, \frac{\|\mathbf{Y}\|^2}{n} \begin{bmatrix} 1 - \frac{\beta_j^2}{\|\beta\|^2+\sigma^2} & -\frac{\beta_j\beta_k}{\|\beta\|^2+\sigma^2} & 0 & 0 \\ -\frac{\beta_j\beta_k}{\|\beta\|^2+\sigma^2} & 1 - \frac{\beta_k^2}{\|\beta\|^2+\sigma^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right),$$

which converges in distribution to $\mathcal{N}(0, (\sigma^2 + \kappa \mathbb{E}[B_0^2]) I_4)$.

□

Theorem 12. *In Setting 2 with the knockoff procedure that takes $\tilde{\mathbf{X}}$ to be an i.i.d. copy of $\mathbf{X}$, let W_j be $f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda)$ for a continuous antisymmetric function f that is not almost everywhere 0, where $\hat{\beta}^\lambda$ is the lasso estimate with penalty parameter λ. Assume α_λ and τ_λ are defined as in Bayati and Montanari (2011) (note that the number of covariates is $2p$ instead of p). Let $G^{(0)}$ be the CDF of $f(\eta(\tau_\lambda Z_1; \alpha_\lambda \tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda \tau_\lambda))$ and $G^{(1)}$ be the CDF of $f(\eta(B_0^{\text{alt}} + \tau_\lambda Z_1; \alpha_\lambda \tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda \tau_\lambda))$, where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independent of B_0^{alt}, and B_0^{alt} has the same distribution as $(B_0 \mid B_0 \neq 0)$. Assume f is such that $G^{(0)}$ and $G^{(1)}$ are CDFs that only have a point mass at 0 and have continuous densities elsewhere. Let (min over an empty set is defined to be infinity)*

$$w_{\text{KF}} = \min\{w \geq 0 : g(w) \leq q\},$$

where

$$g(w) = \begin{cases} \dfrac{\gamma G^{(0)}(-w) + (1-\gamma)(G^{(1)}(-w))}{\gamma(1 - G^{(0)}(w)) + (1-\gamma)(1 - G^{(1)}(w))}, & w > 0, \\ \lim\limits_{w \to 0^+} g(w), & w = 0. \end{cases}$$

For almost every $q \in (0,1)$, one of the following cases is true:

(a) $w_{\text{KF}} > 0$ and $g'(w_{\text{KF}}) \neq 0$, then the FDP and realized power of the knockoffs procedure at level q converge in probability to

$$\frac{\gamma G^{(0)}(-w_{\text{KF}})}{\gamma(1 - G^{(0)}(w_{\text{KF}})) + (1-\gamma)(1 - G^{(1)}(w_{\text{KF}}))} \quad \text{and} \quad 1 - G^{(1)}(w_{\text{KF}}),$$

respectively;

(b) $w_{\text{KF}} = 0$ and either the right derivative $g'(0) < 0$ or $g(0) < q$, then the FDP and realized power of the knockoffs procedure at level q converge in probability to

$$\lim_{w \to 0^+} \frac{\gamma G^{(0)}(-w)}{\gamma(1 - G^{(0)}(w)) + (1-\gamma)(1 - G^{(1)}(w))} \quad \text{and} \quad \lim_{w \to 0^+} 1 - G^{(1)}(w),$$

respectively;

(c) $w_{\mathrm{KF}} = \infty$*, then the realized power of the knockoffs procedure at level* q *converges in probability to* 0.

Proof of Theorem 12. Let

$$\hat{w}_p = \min\{w > 0 : g_p(w) \equiv \frac{1/p + \#\{j : f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \le -w\}/p}{\#\{j : f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \ge w\}/p} \le q\}.^9$$

Similar to the proof of Theorem 11, if we can show the convergence of the empirical CDFs, we can show $\hat{w}_p$ converges in probability to w_{KF} and the results of the theorem then follow. Hence, we only need the results from Lemma 15. □

Lemma 15. *Under the setting in Theorem 12, for any nonzero* $t \in \mathbb{R}$*,*

$$\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \le t) \xrightarrow{p} \mathbb{P}\left(f(\eta(B_0 + \tau_\lambda Z_1; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda\tau_\lambda)) \le t\right)$$

and

$$\frac{1}{\#\{j \in [p] : \beta_j = 0\}}\sum_{j=1}^{p} \mathbb{I}(f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \le t, \beta_j = 0) \xrightarrow{p} \mathbb{P}\left(f(\eta(\tau_\lambda Z_1; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda\tau_\lambda)) \le t\right),$$

where $B_0 \sim \gamma\delta_0 + (1-\gamma)\pi_1$ *and* $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ *are independent of* B_0*. These imply*

$$\frac{1}{\#\{j \in [p] : \beta_j \neq 0\}}\sum_{j=1}^{p} \mathbb{I}(f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \le t, \beta_j \neq 0) \xrightarrow{p} \mathbb{P}\left(f(\eta(B_0^{\mathrm{alt}} + \tau_\lambda Z_1; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda\tau_\lambda)) \le t\right).$$

Proof of Lemma 15. To use the results in Bayati and Montanari (2011), we apply the following re-normalization to Setting 2: assume $\mathbf{X}$ is divided by $\sqrt{n}$ and β is multiplied by $\sqrt{n}$.

For simplicity of notation, we relabel the covariates, so all odd-labeled covariates correspond to real covariates, and all even-labeled covariates correspond to knockoffs. We condition on $\beta_{1:\infty}$. Note that the relabeling means only odd β_j's correspond to draws from $\gamma\delta_0 + (1-\gamma)\pi_1$, and the even β_j's are just zero.

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(f(\hat{\beta}_{2j-1}^\lambda, \hat{\beta}_{2j}^\lambda) \le t)\right] \\
&= \mathbb{P}(f(\hat{\beta}_{2J-1}^\lambda, \hat{\beta}_{2J}^\lambda) \le t) \qquad (J \sim \mathrm{Unif}([p])) \\
&= \mathbb{P}(f(\hat{\beta}_{2J-1}^\lambda, \hat{\beta}_{2J'}^\lambda) \le t) \qquad (J, J' \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([p]), \text{ by exchangeability}).
\end{aligned}$$

We know the limit of this probability if we can show that

$$(\hat{\beta}_{2J-1}^\lambda, \hat{\beta}_{2J'}^\lambda) \xrightarrow{d} (\eta(B_0 + \tau_\lambda Z_1; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda\tau_\lambda)),$$

[9] Formally, we only take the minimum over $w \in \{|f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda)| : f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda) \neq 0\}$. Such a difference is important when $g_p(w) \le q$ for all $w > 0$. The term $1/p$ does not matter asymptotically, in that we can still consider the numerator as an empirical CDF of the $f(\sqrt{n}\hat{\beta}_j^\lambda, \sqrt{n}\hat{\beta}_{j+p}^\lambda)$'s.

where $Z_1, Z_2 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$, independent of B_0.

$$\mathbb{P}(\hat{\beta}^{\lambda}_{2J-1} \le s, \hat{\beta}^{\lambda}_{2J'} \le t) = \mathbb{E}[\hat{F}^{\text{odd}}_p(s)\hat{F}^{\text{even}}_p(t)],$$

where

$$\hat{F}^{\text{odd}}_p(s) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}^{\lambda}_{2j-1} \le s), \hat{F}^{\text{even}}_p(t) = \frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}^{\lambda}_{2j} \le t).$$

We need these two terms to converge in probability, which would give us asymptotic independence of $(\hat{\beta}^{\lambda}_{2J-1}, \hat{\beta}^{\lambda}_{2J'})$ by convergence of CDF via the bounded convergence theorem. Note that we only have to analyze $t \neq 0$, which corresponds to the continuity points. By exchangeability,

$$\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}^{\lambda}_{2j} \le t) \mid \sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0) \sim \text{Hypergeometric}(p + p\gamma_p, \sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0), p).$$

Here,

$$\gamma_p = \frac{\#\{j \in [p] : \beta_{2j-1} = 0\}}{p} \to \gamma.$$

The hypergeometric distribution divided by p has mean

$$\frac{\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0)}{p + p\gamma_p}$$

and variance

$$\frac{\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0)}{p + p\gamma_p}\left(1 - \frac{\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0)}{p + p\gamma_p}\right)\frac{\gamma_p}{p-1}.$$

Lemma 16. *In Setting 2 with the knockoff procedure that takes $\tilde{\mathbf{X}}$ to be an i.i.d. copy of $\mathbf{X}$, assume the ε_i's, X_{ij}'s and $\tilde{X}_{ij}$'s do not change with n, p as long as $n \ge i$ and $p \neq j$, then*

$$\frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}^{\lambda}_{j} \le t, \beta_j = 0) \overset{\text{a.s.}}{\to} \frac{\gamma+1}{2}\mathbb{P}(\eta(\tau_\lambda Z; \alpha_\lambda \tau_\lambda) \le t)$$

for any $t \neq 0$, where $Z \sim \mathcal{N}(0,1)$.

Proof of Lemma 16. For $\varepsilon > 0$, let $\phi(x,y) = \mathbb{I}(x \le t, |y| \le \varepsilon)$. Take

$$\phi_{1,k}(x,y) = 1 - \min(1, k \times \inf_{z \le t \text{ and } |w| \le \varepsilon} \|(z,w) - (x,y)\|),$$

and

$$\phi_{2,k}(x,y) = \min(1, k \times \inf_{z \ge t \text{ or } |w| \ge \varepsilon} \|(z,w) - (x,y)\|).$$

It is clear that

$$\phi_{2,k}(x,y) \le \phi(x,y) \le \phi_{1,k}(x,y),$$

and for each $k > 0$, $\phi_{1,k}$ and $\phi_{2,k}$ are uniformly continuous, so we have almost surely (Bayati and Montanari 2011),

$$\lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \psi(\hat{\beta}^{\lambda}_{j}, \beta_j) = \mathbb{E}[\psi(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}})], \qquad \psi = \phi_{1,k}, \phi_{2,k}, \tag{31}$$

where $Z \sim N(0,1)$ independent of $B_{\text{all}} \sim \frac{\gamma+1}{2}\delta_0 + \frac{1-\gamma}{2}\pi_1$.

Now we assume ε is such that B_{all} does not have point mass at ε, which holds for almost every $\varepsilon > 0$, then

$$
\begin{aligned}
&\mathbb{E}[\phi_{1,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}})] \\
&= \mathbb{P}(B_{\text{all}} = 0)\mathbb{E}[\phi_{1,k}(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda), 0)] + \mathbb{P}(0 < |B_{\text{all}}| \le \varepsilon + \frac{1}{k})\mathbb{E}[\phi_{1,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}}) \mid 0 < |B_{\text{all}}| \le \varepsilon + \frac{1}{k}] \\
&\to \frac{1+\gamma}{2}\mathbb{E}[\phi(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda), 0)] + \mathbb{P}(0 < |B_{\text{all}}| \le \varepsilon)\mathbb{E}[\phi(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}}) \mid 0 < |B_{\text{all}}| \le \varepsilon] \\
&\mathbb{E}[\phi_{2,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}})] \\
&= \mathbb{P}(B_{\text{all}} = 0)\mathbb{E}[\phi_{2,k}(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda), 0)] + \mathbb{P}(0 < |B_{\text{all}}| \le \varepsilon)\mathbb{E}[\phi_{2,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}}) \mid 0 < |B_{\text{all}}| \le \varepsilon] \\
&\to \frac{1+\gamma}{2}\mathbb{E}[\phi(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda), 0)] + \mathbb{P}(0 < |B_{\text{all}}| \le \varepsilon)\mathbb{E}[\phi(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}}) \mid 0 < |B_{\text{all}}| \le \varepsilon]
\end{aligned} \tag{32}
$$

as $k \to \infty$, by the bounded convergence theorem. Since

$$
\frac{1}{2p}\sum_{j=1}^{2p} \phi_{2,k}(\hat{\beta}_j^\lambda, \beta_j) \le \frac{1}{2p}\sum_{j=1}^{2p} \phi(\hat{\beta}_j^\lambda, \beta_j) \le \frac{1}{2p}\sum_{j=1}^{2p} \phi_{1,k}(\hat{\beta}_j^\lambda, \beta_j),
$$

we have,

$$
\limsup_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \phi(\hat{\beta}_j^\lambda, \beta_j) \le \lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \phi_{1,k}(\hat{\beta}_j^\lambda, \beta_j) = \mathbb{E}[\phi_{1,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}})]
$$

and

$$
\liminf_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \phi(\hat{\beta}_j^\lambda, \beta_j) \ge \lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \phi_{2,k}(\hat{\beta}_j^\lambda, \beta_j) = \mathbb{E}[\phi_{2,k}(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}})].
$$

Then it follows from equations (31) and (32) that

$$
\begin{aligned}
&\frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \le t, |\beta_j| \le \varepsilon) \\
&\to \frac{\gamma+1}{2}\mathbb{E}[\phi(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda), 0)] + \mathbb{P}(0 < |B_{\text{all}}| \le \varepsilon)\mathbb{E}[\phi(\eta(B_{\text{all}} + \tau_\lambda Z; \alpha_\lambda\tau_\lambda), B_{\text{all}}) \mid 0 < |B_{\text{all}}| \le \varepsilon].
\end{aligned}
$$

The second term goes to 0 as $\varepsilon \to 0$ since $|\phi| \le 1$. Now we want to show

$$
\lim_{\varepsilon\to 0}\lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \le t, |\beta_j| \le \varepsilon) = \lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \le t, |\beta_j| = 0),
$$

while the difference is

$$
\lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \le t, 0 < |\beta_j| \le \varepsilon) \le \lim_{p\to\infty} \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(0 < |\beta_j| \le \varepsilon) = \frac{1-\gamma}{2}\pi_1((0,\varepsilon]),
$$

which converges to 0 as $\varepsilon \to 0$.

Now we have shown for any $t \neq 0$,

$$\frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \leq t, |\beta_j| = 0) \overset{\text{a.s.}}{\to} \frac{\gamma+1}{2}\mathbb{P}(\eta(\tau_\lambda Z; \alpha_\lambda \tau_\lambda) \leq t).$$

□

By Lemma 16, letting $B_p = \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \leq t, \beta_j = 0)$,

$$\mathbb{E}\left[\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_{2j}^\lambda \leq t)\right] = \frac{2}{1+\gamma_p}\mathbb{E}[B_p] \to \mathbb{P}(\eta(\tau_\lambda Z; \alpha_\lambda \tau_\lambda) \leq t)$$

by the bounded convergence theorem.

$$\begin{aligned}\operatorname{Var}\left[\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_{2j}^\lambda \leq t)\right] &= \mathbb{E}\left[\frac{2B_p}{1+\gamma_p}(1 - \frac{2B_p}{1+\gamma_p})\frac{\gamma_p}{p-1}\right] + \operatorname{Var}\left[\frac{2B_p}{1+\gamma_p}\right] \\ &\leq \frac{\gamma_p}{p-1} + \frac{4}{(1+\gamma_p)^2}\operatorname{Var}[B_p] \to 0,\end{aligned}$$

where $\operatorname{Var}[B_p] \to 0$ by the bounded convergence theorem since B_p converges to a constant. Now we have shown for $t \neq 0$,

$$\hat{F}_p^{\text{even}}(t) \overset{p}{\to} \mathbb{P}(\eta(\tau_\lambda Z; \alpha_\lambda \tau_\lambda) \leq t),$$

and convergence of $\hat{F}_p^{\text{odd}}(s)$ follows from

$$\hat{F}_p^{\text{odd}}(s) = 2 \times \frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \leq s) - \hat{F}_p^{\text{even}}(s),$$

where the convergence of $\frac{1}{2p}\sum_{j=1}^{2p} \mathbb{I}(\hat{\beta}_j^\lambda \leq s)$ for $s \neq 0$ can be established using uniformly continuous functions as upper and lower bounds on the indicator function by the same technique as in Lemma 16.

We have now proved the first result of Lemma 15 in expectation, and proceed to the variance. We need to analyze the following to apply the Markov inequality.

$$\begin{aligned}\mathbb{E}[(\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(f(\hat{\beta}_{2j-1}^\lambda, \hat{\beta}_{2j}^\lambda) \leq t))^2] &= \frac{1}{p}\mathbb{P}(f(\hat{\beta}_{2J_1-1}^\lambda, \hat{\beta}_{2J_1'}^\lambda) \leq t) \\ &\quad + \frac{p-1}{p}\mathbb{P}(f(\hat{\beta}_{2J_1-1}^\lambda, \hat{\beta}_{2J_1'}^\lambda) \leq t, f(\hat{\beta}_{2J_2-1}^\lambda, \hat{\beta}_{2J_2'}^\lambda) \leq t),\end{aligned}$$

where

$$(J_1, J_1', J_2, J_2') \overset{\text{i.i.d.}}{\sim} \operatorname{Unif}(\{(i,j,k,\ell) \in [p]^4 : i \neq k, j \neq \ell\}).$$

We can evaluate this by showing that asymptotically $(\hat{\beta}_{2J_1-1}^\lambda, \hat{\beta}_{2J_1'}^\lambda, \hat{\beta}_{2J_2-1}^\lambda, \hat{\beta}_{2J_2'}^\lambda)$ converges in distribution to four independent random variables. Per the results we have shown, we can define CDFs F_{even} and F_{odd} such that for $t \neq 0$, $\hat{F}_p^{\text{even}}(t) \overset{p}{\to} F_{\text{even}}(t)$ and $\hat{F}_p^{\text{odd}}(t) \overset{p}{\to} F_{\text{odd}}(t)$. Thus,

$$\begin{aligned}\mathbb{P}(\hat{\beta}_{2J_1-1}^\lambda \leq s, \hat{\beta}_{2J_1'}^\lambda \leq t, \hat{\beta}_{2J_2-1}^\lambda \leq s', \hat{\beta}_{2J_2'}^\lambda \leq t') &= \mathbb{E}\left[\hat{F}_{\text{odd}}(s)\hat{F}_{\text{even}}(t)\left(\frac{p\hat{F}_{\text{odd}}(s') - 1}{p-1}\right)\left(\frac{p\hat{F}_{\text{even}}(t') - 1}{p-1}\right)\right] \\ &\to F_{\text{odd}}(s)F_{\text{even}}(t)F_{\text{odd}}(s')F_{\text{even}}(t')\end{aligned}$$

by the bounded convergence theorem, assuming $s, s', t, t' \neq 0$ and $s \leq s'$ and $t \leq t'$ without loss of generality. It follows immediately that $(\hat{\beta}^{\lambda}_{2J_1-1}, \hat{\beta}^{\lambda}_{2J_1'}, \hat{\beta}^{\lambda}_{2J_2-1}, \hat{\beta}^{\lambda}_{2J_2'})$ converges in distribution to four independent random variables. We are now able to claim convergence in probability for

$$\frac{1}{p}\sum_{j=1}^{p} \mathbb{I}(f(\hat{\beta}^{\lambda}_{2j-1}, \hat{\beta}^{\lambda}_{2j}) \leq t).$$

For

$$\frac{1}{|\{j \in [p] : \beta_{2j-1} = 0\}|}\sum_{j=1}^{p} \mathbb{I}(f(\hat{\beta}^{\lambda}_{2j-1}, \hat{\beta}^{\lambda}_{2j}) \leq t, \beta_{2j-1} = 0),$$

let $N_p = \{j \in [p] : \beta_{2j-1} = 0\}$. To show its convergence, we use the same technique where we compute its first and second moments by finding the asymptotic distribution of a four-dimensional random vector. Specifically, We just need to show that

$$(\hat{\beta}^{\lambda}_{2J_1-1}, \hat{\beta}^{\lambda}_{2J_1'}, \hat{\beta}^{\lambda}_{2J_2-1}, \hat{\beta}^{\lambda}_{2J_2'}) \xrightarrow{d} (\eta(\tau_\lambda Z_1; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_2; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_3; \alpha_\lambda\tau_\lambda), \eta(\tau_\lambda Z_4; \alpha_\lambda\tau_\lambda)),$$

where

$$(J_1, J_1', J_2, J_2') \sim \text{Unif}(\{(i,j,k,\ell) \in N_p^4 : i \neq k, j \neq \ell\}) \text{ and } Z_1, Z_2, Z_3, Z_4 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Note that

$$\begin{aligned}
&\mathbb{P}(\hat{\beta}^{\lambda}_{2J_1-1} \leq s, \hat{\beta}^{\lambda}_{2J_1'} \leq t, \hat{\beta}^{\lambda}_{2J_2-1} \leq s', \hat{\beta}^{\lambda}_{2J_2'} \leq t') \\
&\qquad = \mathbb{E}\left[\hat{F}_p^{\text{odd null}}(s)\hat{F}_p^{\text{even null}}(t)\left(\frac{p\gamma_p\hat{F}_p^{\text{odd null}}(s') - 1}{p\gamma_p - 1}\right)\left(\frac{p\gamma_p - \hat{F}_p^{\text{even null}}(t')}{p\gamma_p - 1}\right)\right],
\end{aligned}$$

assuming $s, s', t, t' \neq 0$ and $s \leq s'$ and $t \leq t'$ without loss of generality, where

$$\hat{F}_p^{\text{odd null}}(s) = \frac{1}{|N_p|}\sum_{j \in N_p} \mathbb{I}(\hat{\beta}^{\lambda}_{2j-1} \leq s), \qquad \hat{F}_p^{\text{even null}}(t) = \frac{1}{|N_p|}\sum_{j \in N_p} \mathbb{I}(\hat{\beta}^{\lambda}_{2j} \leq t).$$

The result follows if we show that for $t \neq 0$

$$\hat{F}_p^{\text{odd null}}(t), \hat{F}_p^{\text{even null}}(t) \xrightarrow{p} \mathbb{P}(\eta(\tau_\lambda Z; \alpha_\lambda\tau_\lambda) \leq t).$$

This is true because

$$\gamma_p\hat{F}_p^{\text{odd null}}(t) = 2B_p - \hat{F}_{\text{even}}(t),$$

and $\hat{F}_p^{\text{odd null}}(t) \overset{d}{=} \hat{F}_p^{\text{even null}}(t)$ by exchangeability. □

Theorem 8. *Consider using the test statistics $T = n^{-1}\mathbf{X}^\top\mathbf{Y}$ for the CRT, and $T_j = n^{-1}\mathbf{X}_j^\top\mathbf{Y}$ for multiple testing with CRT p-values and knockoffs. Let M^2_{retro} be the asymptotic second moment of the retrospectively collected Y_i, i.e.,*

$$M^2_{\text{retro}} = \frac{\mathbb{E}[Y^2_{\text{raw}}g(Y_{\text{raw}})]}{\mathbb{E}[g(Y_{\text{raw}})]},$$

where $Y_{\text{raw}} \sim \mathcal{N}(0, \sigma^2 + v_Z^2)$ is drawn from the asymptotic distribution of Y without rejection.[10] *Note that in Setting 4, the corresponding v_Z^2 (or $v^2_{X_{\text{-}j}}$) is equal to $\kappa\mathbb{E}[B_0^2]$.*

[10] M_{retro} always exists because $g(y) \in [0,1]$ and is not almost everywhere zero.

1. *In Setting 3, the asymptotic power of the CRT is equal to that of a z-test with standardized effect size*
$$\frac{hM_{\text{retro}}}{v_Z^2+\sigma^2}.$$

2. *In Setting 4, for almost all $q \in (0,1)$, BH or AdaPT at level q applied to CRT p-values using T_j (or $|T_j|$) have one-sided (or two-sided) effective π_μ given by the distribution of $\frac{M_{\text{retro}}}{\sigma^2+\kappa\mathbb{E}[B_0^2]}B_0$ with respect to BH or AdaPT at level q.*

3. *In Setting 4, for almost all $q \in (0,1)$, knockoffs with $\tilde{\mathbf{X}}$ an i.i.d. copy of $\mathbf{X}$, antisymmetric function $f(x,y) = x - y$, test statistic T_j, and level q has one-sided effective π_μ given by the distribution of $\frac{M_{\text{retro}}}{\sqrt{2}(\sigma^2+\kappa\mathbb{E}[B_0^2])}B_0$ with respect to AdaPT at level q.*

Proof of Theorem 8. We prove the three statements in the theorem one by one. In the proof, we rescale T and $\tilde{T}$ by $\sqrt{n}$, as we will make explicit later when needed.

1. The retrospective sampling does not affect how we run the CRT, since the CRT is carried out using the distribution $\tilde{X} \mid Z \sim \mathcal{N}(Z\xi, 1)$. Hence, we should still analyze (7). Same as the rest of the proof of Theorem 1, we only need to show that instead of $\|Y\|/n \xrightarrow{p} \sigma^2 + v_Z^2$, we have $\|\mathbf{Y}\|^2/n \xrightarrow{p} M_{\text{retro}}^2$, which holds because the Y_i's are i.i.d. with all absolute moments satisfying $(k = 0, 1, \dots)$
$$\mathbb{E}[|Y_i|^k] = \frac{\mathbb{E}[|Y_{\text{non-asymptotic}}|^k g(Y_{\text{non-asymptotic}})]}{\mathbb{E}[g(Y_{\text{non-asymptotic}})]} \to \frac{\mathbb{E}[|Y_{\text{raw}}|^k g(Y_{\text{raw}})]}{\mathbb{E}[g(Y_{\text{raw}})]}, \tag{33}$$
where
$$Y_{\text{non-asymptotic}} \sim \mathcal{N}\left(0, \frac{h^2}{n} + (\theta + h\eta/\sqrt{n})^\top \Sigma_Z (\theta + h\eta/\sqrt{n}) + \sigma^2\right).$$
Equation (33) holds because we can write, for $k = 0, 1, \dots,$
$$\begin{aligned}\mathbb{E}[|Y_{\text{non-asymptotic}}|^k g(Y_{\text{non-asymptotic}})] &= \left(\frac{h^2}{n} + (\theta + h\eta/\sqrt{n})^\top \Sigma_Z (\theta + h\eta/\sqrt{n}) + \sigma^2\right)^{k/2} \\ &\times \mathbb{E}\left[|W|^k g\left(\sqrt{\frac{h^2}{n} + (\theta + h\eta/\sqrt{n})^\top \Sigma_Z (\theta + h\eta/\sqrt{n}) + \sigma^2} W\right)\right], W \sim \mathcal{N}(0,1),\end{aligned}$$
use (8) to establish
$$\sqrt{\frac{h^2}{n} + (\theta + h\eta/\sqrt{n})^\top \Sigma_Z (\theta + h\eta/\sqrt{n}) + \sigma^2} W \xrightarrow{d} Y_{\text{raw}},$$
and finally apply the dominated convergence theorem with $|W|^k$ as the dominating function.

2. Applying Lemma 11, we have
$$\mathbf{X}^\top \mathbf{Y} \mid \mathbf{Y}, \beta \sim \mathcal{N}\left(\frac{\|\mathbf{Y}\|^2}{\|\beta\|^2 + \sigma^2}\beta, \|\mathbf{Y}\|^2 \left(I - \frac{1}{\|\beta\|^2 + \sigma^2}\beta\beta^\top\right)\right).$$
Let $T_j = n^{-1/2}\mathbf{X}_j^\top \mathbf{Y}$. For $j \neq k$,

$$\left(\begin{array}{c} T_j - \frac{M_{\text{retro}}^2 \sqrt{n}\beta_j}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \\ T_k - \frac{M_{\text{retro}}^2 \sqrt{n}\beta_k}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \end{array} \right) \mid \mathbf{Y}, \beta$$
$$\sim \mathcal{N}\left(\left(\frac{\|\mathbf{Y}\|^2/n}{\|\beta\|^2+\sigma^2} - \frac{M_{\text{retro}}^2}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \right) \left(\begin{array}{c} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \end{array} \right), \frac{\|\mathbf{Y}\|^2}{n} \left[\begin{array}{cc} 1 - \frac{\beta_j^2}{\|\beta\|^2+\sigma^2} & -\frac{\beta_j \beta_k}{\|\beta\|^2+\sigma^2} \\ -\frac{\beta_j \beta_k}{\|\beta\|^2+\sigma^2} & 1 - \frac{\beta_k^2}{\|\beta\|^2+\sigma^2} \end{array} \right] \right),$$

which converges in distribution to $\mathcal{N}(0, M_{\text{retro}}^2 I_2)$. We get the desired result by dividing both sides by M_{retro}.

3. We have shown in Lemma 14 that

$$\left(\begin{array}{c} \mathbf{X}^\top \mathbf{Y} \\ \tilde{\mathbf{X}}^\top \mathbf{Y} \end{array} \right) \mid \mathbf{Y}, \beta \sim \mathcal{N}\left(\frac{\|\mathbf{Y}\|^2}{\|\beta\|^2+\sigma^2} \left(\begin{array}{c} \beta \\ 0 \end{array} \right), \|\mathbf{Y}\|^2 \left[\begin{array}{cc} I - \frac{1}{\|\beta\|^2+\sigma^2} \beta\beta^\top & 0 \\ 0 & I \end{array} \right] \right).$$

Let $T_j = n^{-1/2} \mathbf{X}_j^\top \mathbf{Y}$, $\tilde{T}_j = n^{-1/2} \tilde{\mathbf{X}}_j^\top \mathbf{Y}$. For $j \neq k$,

$$\left(\begin{array}{c} T_j - \frac{M_{\text{retro}}^2 \sqrt{n}\beta_j}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \\ T_k - \frac{M_{\text{retro}}^2 \sqrt{n}\beta_k}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \\ \tilde{T}_j \\ \tilde{T}_k \end{array} \right) \mid \mathbf{Y}, \beta$$
$$\sim \mathcal{N}\left(\left(\frac{\|\mathbf{Y}\|^2/n}{\|\beta\|^2+\sigma^2} - \frac{M_{\text{retro}}^2}{\kappa \mathbb{E}[B_0^2]+\sigma^2} \right) \left(\begin{array}{c} \sqrt{n}\beta_j \\ \sqrt{n}\beta_k \\ 0 \\ 0 \end{array} \right), \frac{\|\mathbf{Y}\|^2}{n} \left[\begin{array}{cccc} 1 - \frac{\beta_j^2}{\|\beta\|^2+\sigma^2} & -\frac{\beta_j \beta_k}{\|\beta\|^2+\sigma^2} & 0 & 0 \\ -\frac{\beta_j \beta_k}{\|\beta\|^2+\sigma^2} & 1 - \frac{\beta_k^2}{\|\beta\|^2+\sigma^2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array} \right] \right),$$

which converges in distribution to $\mathcal{N}(0, M_{\text{retro}}^2 I_4)$. We get the desired result by dividing both sides by M_{retro}.

□

F Fixed-X test with the OLS coefficient

Consider the fixed-X test in Section 2.2.2. Under the sequence of alternatives $\beta = h/\sqrt{n}$, the power is

$$\mathbb{P}_{\beta=h/\sqrt{n}}\left(\hat{\beta}_1 > z_\alpha \sigma \sqrt{\hat{\Omega}_{11}} \right) = \mathbb{E}\left[\mathbb{P}_{\beta=h/\sqrt{n}}\left(\hat{\beta}_1 > z_\alpha \sigma \sqrt{\hat{\Omega}_{11}} \mid \mathbf{X}, \mathbf{Z} \right) \right] = \mathbb{E}\left[\Phi\left(\frac{h}{\sigma \sqrt{n \hat{\Omega}_{11}}} - z_\alpha \right) \right].$$

Since $\hat{\Omega} \sim$ Inv-Wishart$(\Omega = \Sigma^{-1}, n)$, where Σ is the joint covariance matrix of (X, Z), we have $n\hat{\Omega}_{11} \xrightarrow{p} \Omega_{11}/(1-\kappa)$ by moment cacluations. Note that $\Omega_{11} = 1$, and it then follows that the asymptotic power of the fixed-X test with OLS coefficient is

$$\mathbb{P}_{\beta=h/\sqrt{n}}(\hat{\beta}_1 > z_\alpha \sigma \sqrt{\hat{\Omega}_{11}}) \to \Phi\left(\frac{h}{\sigma} \sqrt{\frac{1-\kappa}{\Omega_{11}}} - z_\alpha \right) = \Phi\left(\frac{h}{\sigma} \sqrt{1-\kappa} - z_\alpha \right),$$

the same as CRT with the OLS coefficient!

G The CRT with unlabeled data

We discuss the conditional CRT in detail using the concrete example in Section 2.3, i.e., Setting 1 but with the following changes: ξ is unknown; $\mathrm{Var}(X \mid Z) = 1$ but is unknown to the CRT.

We notice that we could write

$$\mathbf{X}_* = \mathbf{Z}_* \xi + \varepsilon_*^X, \varepsilon_*^X \sim \mathcal{N}(0, I_{n_*}),$$

where ε_*^X is independent of $\mathbf{Z}_*$ and ε. It can be seen that under the null hypothesis $H_0 : X \perp\!\!\!\perp Y \mid Z$, we have $\varepsilon_*^X \perp\!\!\!\perp \mathbf{Y} \mid \mathbf{Z}_*$. Then

$$\begin{aligned}
\mathbf{X}_* &= \left(\mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* + \left(I_{n_*} - \mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* \\
&= \left(\mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* + \left(I_{n_*} - \mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \varepsilon_*^X \\
&= \left(\mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* + A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top \varepsilon_*^X \\
&= \left(\mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* + A_{\mathbf{Z}_*} \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\| \frac{A_{\mathbf{Z}_*}^\top \varepsilon_*^X}{\|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|},
\end{aligned}$$

where $A_{\mathbf{Z}_*}$ is an $n_* \times (n_* - p)$ matrix that satisfies

$$A_{\mathbf{Z}_*} A_{\mathbf{Z}_*}^\top = I_{n_*} - \mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top, \quad A_{\mathbf{Z}_*}^\top A_{\mathbf{Z}_*} = I_{n_*-p}.$$

Such an $A_{\mathbf{Z}_*}$ exists because $I_{n_*} - \mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top$ is a projection matrix. Hence, $A_{\mathbf{Z}_*}^\top \varepsilon_*^X \mid \mathbf{Z}_* \sim \mathcal{N}(0, I_{n_*-p})$. Let $H_{\mathbf{Z}_*} = \mathbf{Z}_* \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top$ and it follows that under the null,

$$\mathbf{X}_* \stackrel{d}{=} H_{\mathbf{Z}_*} \mathbf{X}_* + A_{\mathbf{Z}_*} \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\| \cdot \tilde{U} \;\Big|\; \mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*} \mathbf{X}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|,$$

where $\mathcal{L}\left(\tilde{U} \mid \mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*} \mathbf{X}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|\right)$ is the uniform distribution on the sphere $\mathbb{S}^{n_*-p-1}$. Armed with this observation, we now know under the null,

$$T(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*) \stackrel{d}{=} T\left(H_{\mathbf{Z}_*} \mathbf{X}_* + A_{\mathbf{Z}_*} \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\| \cdot \tilde{U}, \mathbf{Y}, \mathbf{Z}_*\right) \;\Big|\; \mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*} \mathbf{X}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|.$$

We know exactly the conditional distribution on the right hand side (at least, we can sample from it to get an empirical estimate), and a cutoff can thus be obtained.

As a concrete example, consider the marginal correlation $T_{\mathrm{MC}}(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*) = \mathbf{Y}^\top \mathbf{X}$. Same as the previous derivation, we write

$$T_{\mathrm{MC}}(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*) = \mathbf{Y}^\top \left(\mathbf{Z} \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_* + \mathbf{Y}^\top \left(\mathbf{X} - \left(\mathbf{Z} \left(\mathbf{Z}_*^\top \mathbf{Z}_*\right)^{-1} \mathbf{Z}_*^\top\right) \mathbf{X}_*\right).$$

The first term is a discardable constant conditional on $\mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*} \mathbf{X}_*, \|A_{\mathbf{Z}_*}^\top \varepsilon_*^X\|$, making the second term the essential part, which admits an interesting interpretation as a generalization of the OLS coefficient with unlabeled data. To elaborate, note that we can write

$$\beta \mathbf{X} + \mathbf{Z}\theta = \beta(I - \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top) \mathbf{X} + \mathbf{Z}(\beta(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X} + \theta)),$$

and thus run OLS equivalently by solving

$$(\hat{\beta}, \hat{\theta}) = \underset{\beta,\theta}{\operatorname{argmin}} \|\mathbf{Y} - \beta(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{X} + \mathbf{Z}(-\beta(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X} - \theta))\|_2^2.$$

Then we have

$$\hat{\beta} = [\mathbf{X}^\top(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)^2\mathbf{X}]^{-1}\mathbf{X}^\top(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{Y} = \frac{\sum_{i=1}^n \tilde{\tau}_i Y_i}{\sum_{i=1}^n \tilde{\tau}_i^2}, \tag{34}$$

where

$$\tilde{\tau} = \mathbf{X} - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{X} \mid \mathbf{Z} \sim \mathcal{N}(0, I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top).$$

Since the denominator of (34) satisfies $n^{-1}\sum_{i=1}^n \tilde{\tau}_i^2 \xrightarrow{p} (1-\kappa)$, the essence of OLS statistic without unlabeled data is $\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{X}$. A natural generalization is thus $\mathbf{Y}^\top\left(I_{n\times n_*} - \mathbf{Z}\left(\mathbf{Z}_*^\top\mathbf{Z}_*\right)^{-1}\mathbf{Z}_*^\top\right)\mathbf{X}_*$.

On the other hand, if we consider the original OLS statistic, its essence $\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\mathbf{X}$ is equal to $\mathbf{Y}^\top(I - \mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top)\varepsilon^X$, and the only parameter unknown for its distribution given $\mathbf{Y}, \mathbf{Z}_*$ is the scalar $\mathrm{Var}(X \mid Z)$. As we eventually learn $\mathrm{Var}(X \mid Z)$ asymptotically, even from the labeled data alone, it makes no difference by conditioning on the unlabeled data.

Now we proceed by removing the constant from T_{MC},

$$\begin{aligned} T_{\mathrm{MC}}^{\mathrm{ess}}(\mathbf{X}_*, \mathbf{Y}, \mathbf{Z}_*) &= \mathbf{Y}^\top\left(\mathbf{X} - \left(\mathbf{Z}\left(\mathbf{Z}_*^\top\mathbf{Z}_*\right)^{-1}\mathbf{Z}_*^\top\right)\mathbf{X}_*\right) \\ &= \mathbf{Y}^\top\left(I_{n\times n_*} - \mathbf{Z}\left(\mathbf{Z}_*^\top\mathbf{Z}_*\right)^{-1}\mathbf{Z}_*^\top\right)\varepsilon_*^X \overset{d}{=} \mathbf{Y}^\top I_{n\times n_*} A_{\mathbf{Z}_*}\|A_{\mathbf{Z}_*}^\top\varepsilon_*^X\| \cdot \tilde{U} \;\Big|\; \mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*}\mathbf{X}_*, \|A_{\mathbf{Z}_*}^\top\varepsilon_*^X\|. \end{aligned}$$

The upper quantile of the last distribution could be obtained by Monte Carlo simulations. We notice that if we know $\mathrm{Var}(X \mid Z) = 1$, we do not have to condition on $\|A_{\mathbf{Z}_*}^\top\varepsilon_*^X\|$ and we can directly use the quantile of $\mathbf{Y}^\top\left(I_{n\times n_*} - \mathbf{Z}\left(\mathbf{Z}_*^\top\mathbf{Z}_*\right)^{-1}\mathbf{Z}_*^\top\right)\varepsilon_*^X$, which simply follows an explicit Gaussian distribution conditional on $\mathbf{Z}_*, \mathbf{Y}, H_{\mathbf{Z}_*}\mathbf{X}_*$. In fact, in our power analysis, we first make the above observation formal and show that we can indeed assume $\mathrm{Var}(X \mid Z)$ is known.

H Comparison of the two τ_λ's for the CRT and knockoffs

For presentational simplicity, we write $\tau_{\lambda_{\mathrm{CRT}}}$ and $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$ for $\tau_{\lambda_{\mathrm{CRT}}}^{\mathrm{CRT}}$ and $\tau_{\lambda_{\mathrm{KF}}}^{\mathrm{KF}}$. We would like to show that the lowest $\tau_{\lambda_{\mathrm{CRT}}}$ is smaller than the lowest $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$. Before we start, we make an important note on the notation: there are three parameters α, τ, and λ in Bayati and Montanari (2011) to characterize the lasso asymptotics, and τ and λ are defined as functions of α over a suitable range. In the rest of this article, we add subscript λ to α and τ (α_λ and τ_λ) to indicate their association with λ. In this section, we need to consider the change of τ as α varies, and as mentioned at the beginning of this section, there are two different τ's for CRT and knockoffs, so we use α without a subscript as a variable that varies, $\tau_{\lambda_{\mathrm{CRT}}}$ and $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$ as functions of α in CRT and knockoffs settings, respectively, and τ as a dummy variable that does not depend on α. We emphasize that the α in this section is purely an AMP parameter and has nothing to do the level of a hypothesis test.

For an α in a suitable range, we have $\tau_{\lambda_{\mathrm{CRT}}}$ and $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$ as implicit functions of α via equations

$$\begin{aligned} \tau_{\lambda_{\mathrm{CRT}}}^2 &= \sigma^2 + \kappa\mathbb{E}[(\eta(B_0 + \tau_{\lambda_{\mathrm{CRT}}}Z; \alpha\tau_{\lambda_{\mathrm{CRT}}}) - B_0)^2], \\ \tilde{\tau}_{\lambda_{\mathrm{KF}}}^2 &= \sigma^2 + 2\kappa\mathbb{E}[(\eta(IB_0 + \tilde{\tau}_{\lambda_{\mathrm{KF}}}Z; \alpha\tilde{\tau}_{\lambda_{\mathrm{KF}}}) - IB_0)^2], \end{aligned} \tag{35}$$

where I is a Bernoulli random variable with parameter $1/2$ independent of B_0 and Z. By Proposition 1.3 in Bayati and Montanari (2011), every valid α for the knockoff setting is a valid α for the CRT setting (since the left hand side of Equation (1.14) is non-increasing in α and the CRT setting doubles the δ parameter there compared to knockoffs). Therefore, if we can show that for every α valid in the knockoff setting, it defines a $\tau_{\lambda_{\mathrm{CRT}}}$ smaller than $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$, which is valid and thus actually corresponds to a λ, then we have shown that the lowest $\tau_{\lambda_{\mathrm{CRT}}}$ is smaller than the lowest $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$.

From now on, we fix α and thus fix $\tau_{\lambda_{\mathrm{CRT}}}$ and $\tilde{\tau}_{\lambda_{\mathrm{KF}}}$ as functions of α. We first see that for any τ,

$$\mathbb{E}[(\eta(B_0+\tau Z;\alpha\tau)-B_0)^2] = \gamma\mathbb{E}[\eta(\tau Z;\alpha\tau)^2] + (1-\gamma)\mathbb{E}_{B\sim\pi_1}[(\eta(B+\tau Z;\alpha\tau)-B^2)],$$

and

$$\mathbb{E}[(\eta(IB_0+\tau Z;\alpha\tau)-IB_0)^2] = \frac{1+\gamma}{2}\mathbb{E}[\eta(\tau Z;\alpha\tau)^2] + \frac{1-\gamma}{2}\mathbb{E}_{B\sim\pi_1}[(\eta(B+\tau Z;\alpha\tau)-B^2)],$$

so

$$\mathbb{E}[(\eta(B_0+\tilde{\tau}_{\lambda_{\mathrm{KF}}}Z;\alpha\tilde{\tau}_{\lambda_{\mathrm{KF}}})-B_0)^2] < 2\mathbb{E}[(\eta(IB_0+\tilde{\tau}_{\lambda_{\mathrm{KF}}}Z;\alpha\tilde{\tau}_{\lambda_{\mathrm{KF}}})-IB_0)^2].$$

By (35), we immediately have

$$\tilde{\tau}^2_{\lambda_{\mathrm{KF}}} > \sigma^2 + \kappa\mathbb{E}[(\eta(B_0+\tilde{\tau}_{\lambda_{\mathrm{KF}}}Z;\alpha\tilde{\tau}_{\lambda_{\mathrm{KF}}})-B_0)^2].$$

When $\tau^2 \to 0$, we have

$$\tau^2 < \sigma^2 < \sigma^2 + \kappa\mathbb{E}[(\eta(B_0+\tau Z;\alpha\tau)-B_0)^2].$$

Now consider the function $f_\alpha(\tau) = \sigma^2 + \kappa\mathbb{E}[(\eta(B_0+\tau Z;\alpha\tau)-B_0)^2] - \tau^2$, where we have seen $f_\alpha(\tilde{\tau}_{\lambda_{\mathrm{KF}}}) < 0$ and $f_\alpha(0^+) > 0$. We can show f_a is a continuous function of τ on $(0, \tilde{\tau}_{\lambda_{\mathrm{KF}}}]$ with the dominated convergence theorem, because

$$(\eta(B_0+\tau Z;\alpha\tau)-B_0)^2 \le \max\left((\tau Z+\alpha\tau)^2, (\tau Z-\alpha\tau)^2\right) \le \underbrace{\tilde{\tau}^2_{\lambda_{\mathrm{KF}}}\max((Z+\alpha)^2,(Z-\alpha)^2)}_{\text{dominating function}}.$$

By continuity, there is at least one $\tau \in (0, \tilde{\tau}_{\lambda_{\mathrm{KF}}})$ that satisfies $f_a(\tau) = 0$. i.e.,

$$\tau^2 = \sigma^2 + \kappa\mathbb{E}[(\eta(B_0+\tau Z;\alpha\tau)-B_0)^2].$$

Due to uniqueness (Proposition 1.3 in Bayati and Montanari (2011)), this solution is $\tau_{\lambda_{\mathrm{CRT}}}$ and thus $\tau^2_{\lambda_{\mathrm{CRT}}} \in (0, \tilde{\tau}^2_{\lambda_{\mathrm{KF}}})$.

I Simulation details

I.1 Oracle methods

Before presenting our simulation results, we discuss some Bayesian methods as baselines that we compare against. They are referred to as oracle methods, because they require the knowledge of the prior distribution of the parameters, which our methods do not use and we generally do not expect to be available in practice.

I.1.1 Controlling the Bayesian FDR

In the multiple testing problem, when we know the prior distribution of the parameters (e.g., when we know γ and π_1 in Setting 2), we can run a Bayesian procedure to incorporate the prior knowledge that we have. Suppose we have the posterior probabilities of the covariates being non-null as $p_1^B, \ldots, p_p^B$. Without loss of generality, we assume they are ordered from large to small. Then we find k such that

$$\frac{\sum_{j=1}^{k} p_j^B}{k} > 1 - q > \frac{\sum_{j=1}^{k+1} p_j^B}{k+1}$$

and reject $1, 2, \ldots, k$. Finally, we reject $k+1$ with probability r, where r satisfies

$$(1-r)\frac{\sum_{j=1}^{k} p_j^B}{k} + r\frac{\sum_{j=1}^{k+1} p_j^B}{k+1} = 1 - q.$$

It is straightforward to see that this procedure controls the Bayesian FDR at level q. In fact, the above procedure controls

$$\mathbb{E}[\text{FDP} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}] \leq q.$$

This is neither stronger nor weaker than the FDR control conditional on the parameters elsewhere in the article:

$$\mathbb{E}[\text{FDP} \mid \text{parameters of the model}] \leq q,$$

while they both control the unconditional FDR. In the simulations, we run a Gibbs sampler to estimate those posterior probabilities.

I.1.2 Bayesian statistic is optimal for the CRT

Here, we show that in Setting 1, the posterior probability is the optimal statistic to use for the CRT. Suppose we have a true prior π for θ, where π is a mixture of δ_0 and π_1 and π_1 has no point mass at 0. Then by the Neyman–Pearson Lemma, the optimal level-α test for $H_0 : \beta = 0$ against $H_1 : \beta \sim \pi_1$ among valid CRTs (i.e., tests conditional on $\mathbf{Y}, \mathbf{Z}$) is the likelihood ratio test that rejects when

$$T_{\text{opt}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = \frac{\iint \mathbb{P}(\mathbf{X}, \mathbf{Z}) \mathbb{P}_{\theta,\beta}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \pi(\theta)\, d\theta\, \pi_1(\beta)\, d\beta}{\int \mathbb{P}(\mathbf{X}, \mathbf{Z}) \mathbb{P}_{\theta,\beta=0}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \pi(\theta)\, d\theta} > c_\alpha(\mathbf{Y}, \mathbf{Z}),$$

where c_α is an appropriate cutoff (if X is discrete, randomize when $T_{\text{opt}} = c_\alpha$). Interestingly, if we have an almost-correct prior on β, i.e., for a $\gamma \in (0, 1)$, $\beta \sim \gamma\delta_0 + (1-\gamma)\pi_1$, and $\beta \perp\!\!\!\perp \theta$ a priori, then the posterior probability of H_1 is

$$\begin{aligned}
&\mathbb{P}(H_1 \text{ holds} \mid \mathbf{X}, \mathbf{Y}, \mathbf{Z}) \\
&\quad = \frac{(1-\gamma) \iint \mathbb{P}(\mathbf{X}, \mathbf{Z}) \mathbb{P}_{\theta,\beta}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \pi(\theta)\, d\theta\, \pi_1(\beta)\, d\beta}{\gamma \int \mathbb{P}(\mathbf{X}, \mathbf{Z}) \mathbb{P}_{\theta,\beta=0}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \pi(\theta)\, d\theta + (1-\gamma) \iint \mathbb{P}(\mathbf{X}, \mathbf{Z}) \mathbb{P}_{\theta,\beta}(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \pi(\theta)\, d\theta\, \pi_1(\beta)\, d\beta},
\end{aligned}$$

which is a monotone function of $T_{\text{opt}}(\mathbf{X}, \mathbf{Y}, \mathbf{Z})$. Hence, the posterior probability is equivalent to the likelihood ratio and thus optimal, regardless of whether γ is correctly specified.

We point out that the Bayesian methods shown in Figure 7 are not the BH-CRT using the oracle statistic introduced in this section, which is prohibitively expensive to simulate for large n and p, and we expect it to have similar performance to the BH-CRT with distilled lasso based on experiments in Section 5.1.

I.2 Comparison of BH and AdaPT applied to CRT p-values

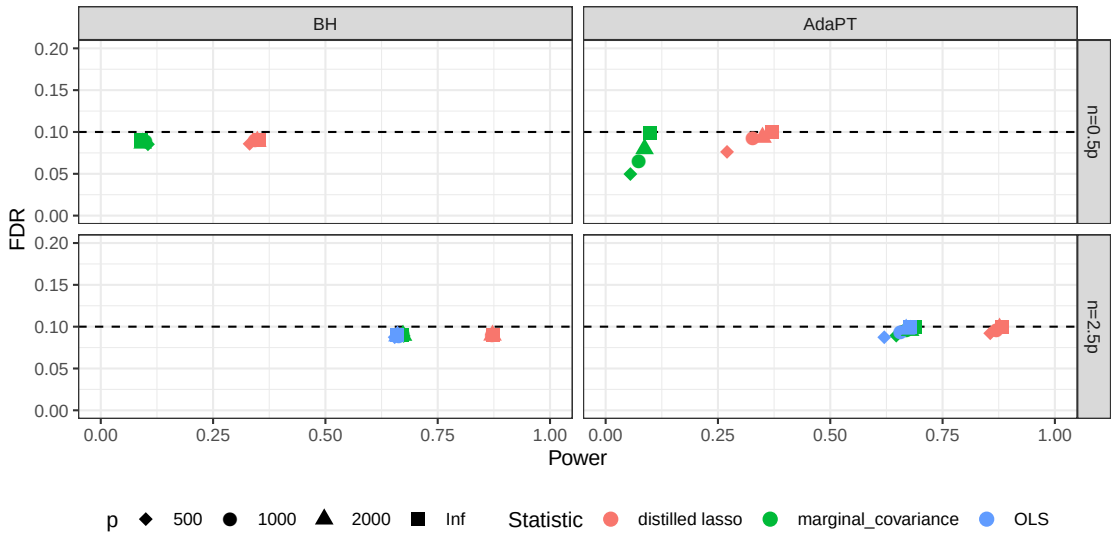

Figure 9: Comparison of BH and AdaPT applied to CRT p-values at FDR level 0.1. The settings are the same as in Figure 7. All standard errors are below 0.01.