Gaussian Universal Likelihood Ratio Testing

BY ROBIN DUNN, AADITYA RAMDAS, SIVARAMAN BALAKRISHNAN, LARRY WASSERMAN
Department of Statistics & Data Science, Carnegie Mellon University,
Pittsburgh, Pennsylvania 15213, U.S.A.
robind@stat.cmu.edu aramdas@stat.cmu.edu siva@stat.cmu.edu larry@stat.cmu.edu

SUMMARY

The likelihood ratio test (LRT) based on the asymptotic chi-squared distribution of the log likelihood is one of the fundamental tools of statistical inference. A recent universal LRT approach based on sample splitting provides valid hypothesis tests and confidence sets in any setting for which we can compute the split likelihood ratio statistic (or, more generally, an upper bound on the null maximum likelihood). The universal LRT is valid in finite samples and without regularity conditions. This test empowers statisticians to construct tests in settings for which no valid hypothesis test previously existed. For the simple but fundamental case of testing the population mean of d-dimensional Gaussian data, the usual LRT itself applies and thus serves as a perfect test bed to compare against the universal LRT. This work presents the first in-depth exploration of the size, power, and relationships between several universal LRT variants. We show that a repeated subsampling approach is the best choice in terms of size and power. We observe reasonable performance even in a high-dimensional setting, where the expected squared radius of the best universal LRT confidence set is approximately 3/2 times the squared radius of the standard LRT-based set. We illustrate the benefits of the universal LRT through testing a non-convex doughnut-shaped null hypothesis, where a universal inference procedure can have higher power than a standard approach.

Some key words: Hypothesis testing; Sample splitting; Universal inference.

1. INTRODUCTION

Suppose we have data from an unknown distribution P_{θ^*} which belongs to some set of distributions $(P_\theta : \theta \in \Theta)$. We wish to test the composite null hypothesis $H_0 : \theta^* \in \Theta_0$. We use the observed data to construct a test statistic T_n and reject H_0 if $T_n > c_\alpha$, where c_α must satisfy

$$\sup_{\theta^* \in \Theta_0} P_{\theta^*}(T_n > c_\alpha) \leq \alpha.$$

Consider, for example, the alternative $H_1 : \theta \in \Theta \setminus \Theta_0$. The generalized likelihood ratio statistic is $\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)$, where $\widehat{\theta}$ is the maximum likelihood estimate (MLE) in Θ and $\widehat{\theta}_0$ is the MLE in Θ_0. We reject H_0 when $2\log\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)\} > c_{\alpha,d}$, where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution and $d = df(\Theta) - df(\Theta_0)$. This construction arises from Wilks' Theorem (Wilks, 1938), which states that $2\log\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)\}$ has an asymptotic χ^2_d distribution under certain regularity conditions. This will apply, for instance, when we have independent and identically distributed (iid) data from an exponential family, Θ_0 is a subset of Θ, and Θ and Θ_0 are linear subspaces in Euclidean space (Van der Vaart, 2000, Theorem 4.6). We can invert the likelihood ratio test (LRT) to produce an asymptotically valid

$100(1-\alpha)\%$ confidence region of the form:

$$C_n^{\mathrm{LRT}}(\alpha) = \left\{\theta \in \Theta : 2\log\left\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\theta)\right\} \le c_{\alpha,d}\right\}.$$

We reject H_0 if and only if $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset$, which is equivalent to rejecting H_0 if and only if $2\log\{\mathcal{L}(\widehat{\theta}) \,/\, \mathcal{L}(\widehat{\theta}_0)\} > c_{\alpha,d}$. We refer to this testing framework as the classical LRT. Some composite nulls are irregular, meaning that Wilks' theorem does not apply and calculating a threshold can be hard due to intractable asymptotics.

The universal inference approach developed by Wasserman et al. (2020) provides a new likelihood ratio testing framework that addresses situations where the classical LRT is not valid. This new LRT relies on sample splitting to construct a test and confidence interval that are valid in finite samples and without regularity conditions. This universal inference method allows one to construct valid tests in settings for which no hypothesis test with type I error control and finite sample guarantees previously existed. The statistical literature has repeatedly emphasized the inadequacy of the asymptotic χ^2 approximation in the small sample setting. Examples include Bartlett (1937), Lehmann (2012), and Medeiros & Ferrari (2017). Small sample sizes also pose a recurrent problem across biological science research. For instance, researchers have noted the prevalence of low-powered studies in neuroscience (Button et al., 2013) and the need for clinical trial designs that account for the small sample sizes common to rare disease and pediatric population research (Ildstad et al., 2001; McMahon et al., 2016).

Many basic questions remain unanswered about the universal LRT, since its power even in very simple settings remains unknown. Further, Wasserman et al. (2020) describe numerous settings in which the universal LRT is the first hypothesis test with finite sample validity. These settings include testing the number of components in mixture models (Hartigan, 1985; McLachlan, 1987; Chen et al., 2009; Li & Chen, 2010) and testing whether the underlying density satisfies the shape constraint of log-concavity (Cule et al., 2010; Axelrod et al., 2019). As a precursor to studying the power in these important but as-yet intractable settings, we first study the universal LRT in the fundamental case of constructing confidence regions (or hypothesis tests) for the population mean $\theta^* \in \mathbb{R}^d$ when $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. In this setting — where the classical LRT is of course valid — our results showcase the reasonable performance of the universal LRT in comparison to the classical approach. With more technical effort, the results can be extended to models that satisfy standard regularity conditions such as quadratic mean differentiability, where the MLE is asymptotically normal (Van der Vaart, 2000, Chapter 5).

This work provides two main contributions: First, we provide a careful analysis of several variants of the universal LRT in the Gaussian case. We show that a repeated subsampling approach is the best choice in terms of size and power. We observe reasonable performance in a high-dimensional setting, where the expected squared radius of the best universal LRT confidence set is approximately 3/2 times the squared radius of the set constructed through the classical approach. Thus, in particular, the power of the universal approaches has the same behavior (in n, d, α) as the classical approach. Second, we show an example of a hypothesis test on normally distributed data where universal LRT methods have higher power than classical testing methods. Specifically, when testing the non-convex "doughnut" null $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$ on $N(\theta^*, I_d)$ data, a universal LRT approach can have higher power than a standard approach that uses the classical LRT confidence set. A test of this form could examine, for instance, whether trial outcomes or biomarker levels are within an acceptable range.

2. Universal LRT confidence sets

2.1. *Universal LRT background*

Wasserman et al. (2020) presented an alternative to the LRT that is valid in finite samples without requiring regularity conditions. Suppose we have n iid observations $Y_1, \ldots, Y_n \sim P_{\theta^*}$, where P_{θ^*} is from a family $(P_\theta : \theta \in \Theta)$. Each P_θ has a density denoted by p_θ. We denote the dataset by $\mathcal{D} = \{Y_1, \ldots, Y_n\}$. To implement the test, first split the data into $\mathcal{D}_0$ and $\mathcal{D}_1$. Let $\widehat{\theta}_1$ be an estimator constructed from $\mathcal{D}_1$. The parameter $\widehat{\theta}_1$ could be the MLE, but any parameter that is fixed given $\mathcal{D}_1$ is valid. Certain choices of $\widehat{\theta}_1$ may be more efficient. Using the data in $\mathcal{D}_0$, the likelihood function is $\mathcal{L}_0(\theta) = \Pi_{Y_i \in \mathcal{D}_0} p_\theta(Y_i)$. Define the split LRT statistic as

$$T_n(\theta) = \mathcal{L}_0(\widehat{\theta}_1)/\mathcal{L}_0(\theta).$$

The universal confidence set for θ^* using the split LRT is

$$C_n^{\text{split}}(\alpha) = \{\theta \in \Theta : T_n(\theta) < 1/\alpha\}.$$

Theorem 1. *$C_n^{split}(\alpha)$ is a valid $100(1-\alpha)\%$ confidence set for θ^*. As a consequence (and equivalently), when testing an arbitrary composite null $H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta \setminus \Theta_0$, rejecting H_0 when $\Theta_0 \cap C_n^{split}(\alpha) = \emptyset$ provides a valid level α hypothesis test. The latter rule reduces to rejecting if $T_n(\widehat{\theta}_0) \geq 1/\alpha$, where $\widehat{\theta}_0 \in \arg\max_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$ is the null MLE.*

Theorem 1 is due to Wasserman et al. (2020). The validity of the universal test does not depend on large samples or regularity conditions. The proof establishes that $E_{\theta^*}\{T_n(\theta^*)\} \leq 1$ and then invokes Markov's inequality. See Section S1 of the supplement for more details. This property on the expectation makes $T_n(\theta^*)$ an e-variable (Grünwald et al., 2020).

The validity of $C_n^{\text{split}}(\alpha)$ only depends on the fact that $\mathbb{E}_{\theta^*}\{T_n(\theta^*)\} \leq 1$. If we consider multiple test statistics that each satisfy this condition, then the average of those test statistics will satisfy the condition as well. Therefore, the average of test statistics $T_n(\theta^*)$ across multiple data splits is also a valid test statistic.

2.2. *Classical test in normal setting*

Assume $Y_1, \ldots, Y_n$ are d-dimensional iid vectors drawn from $N(\theta^*, I_d)$ with $\theta^* \in \mathbb{R}^d$. Where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution, the classical LRT confidence set for θ^* is

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\right\}. \tag{1}$$

See Section S2 of the supplement for a derivation of (1). In this case, $C_n^{\text{LRT}}(\alpha)$ is valid in finite samples, since $n\|\theta^* - \overline{Y}\|^2$ follows a χ^2_d distribution. We compare $C_n^{\text{LRT}}(\alpha)$ to the split LRT set and several universal confidence sets that are variants of the split LRT set.

2.3. *Split, cross-fit, and subsampling sets in normal setting*

First, we consider two universal LRT variants based on a single split of the data. Assume we split the n observations in half, such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $n/2$ observations. Define $\overline{Y}_0 = (2/n)\sum_{Y_i \in \mathcal{D}_0} Y_i$ and $\overline{Y}_1 = (2/n)\sum_{Y_i \in \mathcal{D}_1} Y_i$. Then the confidence set for θ^* based on the split likelihood ratio is

$$\begin{aligned} C_n^{\text{split}}(\alpha) &= \left\{\theta \in \Theta : \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) < \frac{1}{\alpha}\right\} \\ &= \left\{\theta \in \Theta : \|\theta - \overline{Y}_0\|^2 < (4/n)\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}. \end{aligned} \tag{2}$$

See Section S2 of the supplement for a derivation of (2). Using the same split, we define the cross-fit statistic as $S_n(\theta) = \{T_n(\theta) + T_n^{\text{swap}}(\theta)\}/2$, where $T_n^{\text{swap}}(\theta)$ is computed by switching the roles of $\mathcal{D}_0$

and $\mathcal{D}_1$. Then the cross-fit confidence set is a valid $100(1-\alpha)\%$ set given by

$$C_n^{\text{CF}}(\alpha) = \left\{\theta \in \Theta : \frac{1}{2}\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2\right)\left\{\exp\left(\frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) + \exp\left(\frac{n}{4}\|\overline{Y}_1 - \theta\|^2\right)\right\} < \frac{1}{\alpha}\right\}.$$

The split and cross-fit sets have both statistical randomness (due to the random sampling of observations) and algorithmic randomness (due to the randomness in splitting the sample into $\mathcal{D}_0$ and $\mathcal{D}_1$). In contrast, the classical LRT only has statistical randomness, since the test is deterministic for a given set of observations. We now consider a repeated subsampling approach. This universal method attempts to mitigate the algorithmic randomness from the split and cross-fit LRTs by splitting the observations many times and averaging the test statistics. Algorithm 1 shows how to compute the subsampling test statistic $T_n(\theta)$ at a given $\theta \in \mathbb{R}^d$.

Algorithm 1 Compute the subsampling test statistic $T_n(\theta)$.

Input: n independent d-dimensional observations $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$ (θ^* unknown), a value of $\theta \in \mathbb{R}^d$, number of subsamples B.
Output: The subsampling test statistic $T_n(\theta)$.
For $b = 1$ to $b = B$:
  Randomly split the data into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$, each containing $n/2$ values of Y_i.
  Let $\overline{Y}_{0,b} = (2/n)\sum_{Y_i \in \mathcal{D}_{0,b}} Y_i$ and let $\overline{Y}_{1,b} = (2/n)\sum_{Y_i \in \mathcal{D}_{1,b}} Y_i$.
  Compute $T_{n,b}(\theta) = \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right)$.
Output the subsampling test statistic $T_n(\theta) = B^{-1}\sum_{b=1}^{B} T_{n,b}(\theta)$.

As noted earlier, this method is also valid. The $100(1-\alpha)\%$ subsampling confidence set is

$$C_n^{\text{subsplit}}(\alpha) = \left\{\theta \in \Theta : \frac{1}{B}\sum_{b=1}^{B}\exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right) < \frac{1}{\alpha}\right\}.$$

Figure 1 shows coverage regions of the classical LRT, split LRT, cross-fit LRT, and subsampling LRT ($B = 100$) from six simulations with $\theta^* = (0, 0)$. We generate 1000 observations from $N(\theta^*, I_2)$, and we use this sample for all simulations. Hence, the variation in the split, cross-fit, and subsampling LRTs across simulations is due to algorithmic randomness.

The coverage regions in Fig. 1 suggest several relationships that we will formalize. We see that the classical LRT provides the smallest confidence regions. This is not surprising since, even in finite samples, the classical LRT statistic follows a chi square distribution under $H_0 : \theta = \theta^*$ in the Gaussian case. The volume of the cross-fit LRT set is less than or equal to the volume of the split LRT set, although the cross-fit set is not entirely contained within the split set. The split and cross-fit approaches both use a single split of the data, but there is a notable improvement from cross-fitting. The subsampling set also has less volume than the split LRT set. Recall that we construct the subsampling test statistic by performing the split LRT over repeated splits of the data and then averaging the test statistics $T_{n,b}(\theta)$. While any individual split LRT region is guaranteed to be spherical, the subsampling set is not necessarily a spherical region. For large B, however, we see that the subsampling region is approximately spherical. Thus, although the subsampling approach is computationally intensive, this hints that it may be possibly to derive a formulaic approximation to the limiting subsampling set.

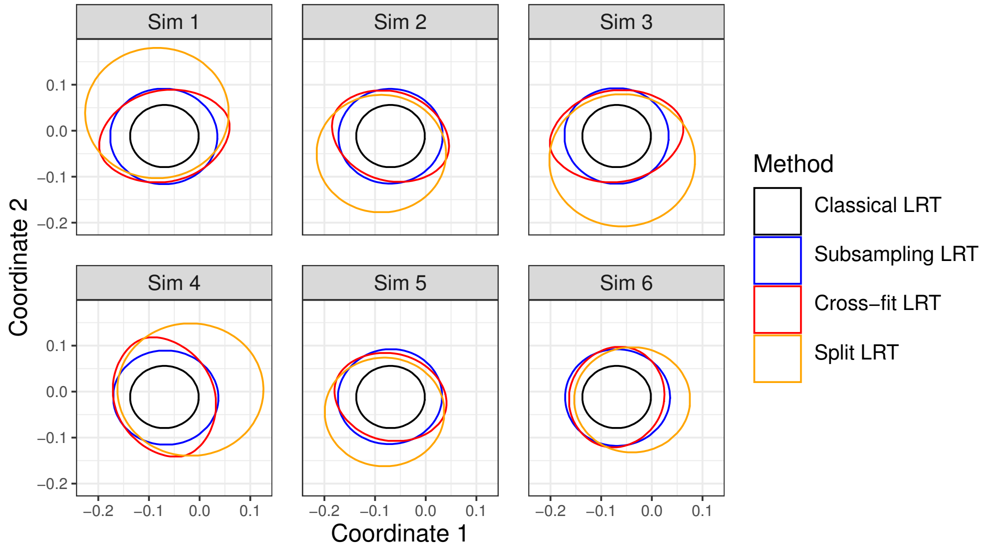

Fig. 1. Coverage regions of classical LRT (black), subsampling LRT (blue), cross-fit LRT (red), and split LRT (orange) at $\alpha = 0.1$. The six simulations use the same 1000 observations from $N(\theta^*, I_2)$ under $\theta^* = (0, 0)$.

2.4. *Limit of subsampling region*

We are particularly interested in the behavior of the subsampling confidence set as $B \to \infty$. Since $B^{-1} \sum_{b=1}^{B} T_{n,b}(\theta) \to \mathbb{E}\{T_n(\theta) \mid \mathcal{D}\}$ as $B \to \infty$, the limiting subsampling set has no algorithmic randomness. We see hints of this in Fig. 1, where the subsampling set at $B = 100$ does not vary much across six simulations on the same data. Theorem 2 describes conditions for the convergence of the ratio of $\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\}$ to an approximation. We have been suppressing the n subscript when it is clear we are working with a single dataset with n observations. Theorem 2 considers a sequence of datasets, so we use the n subscript to index the datasets.

Theorem 2. *Assume we have a sequence of datasets $(\mathcal{D}_n)_{n \in 2\mathbb{N}}$, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$, and let $\mathcal{D}_{1,n} = \mathcal{D}_n \backslash \mathcal{D}_{0,n}$. Define $\overline{Y}_n = (1/n) \sum_{i=1}^{n} Y_{ni}$, $\overline{Y}_{0,n} = (2/n) \sum_{Y_{ni} \in \mathcal{D}_{0,n}} Y_{ni}$, and $\overline{Y}_{1,n} = (2/n) \sum_{Y_{ni} \in \mathcal{D}_{1,n}} Y_{ni}$. Let $c > 0$, and let (θ_n) be a sequence that satisfies $\|\overline{Y}_n - \theta_n\| \le c/\sqrt{n}$ for all n. Then*

$$\mathbb{E}\{T_n(\theta_n) \mid \mathcal{D}_n\} \Big/ \left\{ \exp\left(\frac{3n}{10} \|\overline{Y}_n - \theta_n\|^2 \right) \left(\frac{2}{5} \right)^{d/2} \right\} = 1 + o_P(1). \tag{3}$$

In words, the subsampling statistic is approximately given by $R(\theta)^{3/5}(2/5)^{d/2}$ where $R(\theta) = \mathcal{L}(\widehat{\theta})/\mathcal{L}(\theta)$ is the usual likelihood ratio statistic.

Section S1 of the supplement contains a proof of Theorem 2. The proof relies critically on the finite sample central limit theorems from Hájek (1960) and Li & Ding (2017) and on the Portmanteau Theorem proof techniques from Van der Vaart (2000).

Since

$$\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\} \approx \exp\left(\frac{3n}{10} \|\overline{Y} - \theta\|^2 \right) \left(\frac{2}{5} \right)^{d/2}, \tag{4}$$

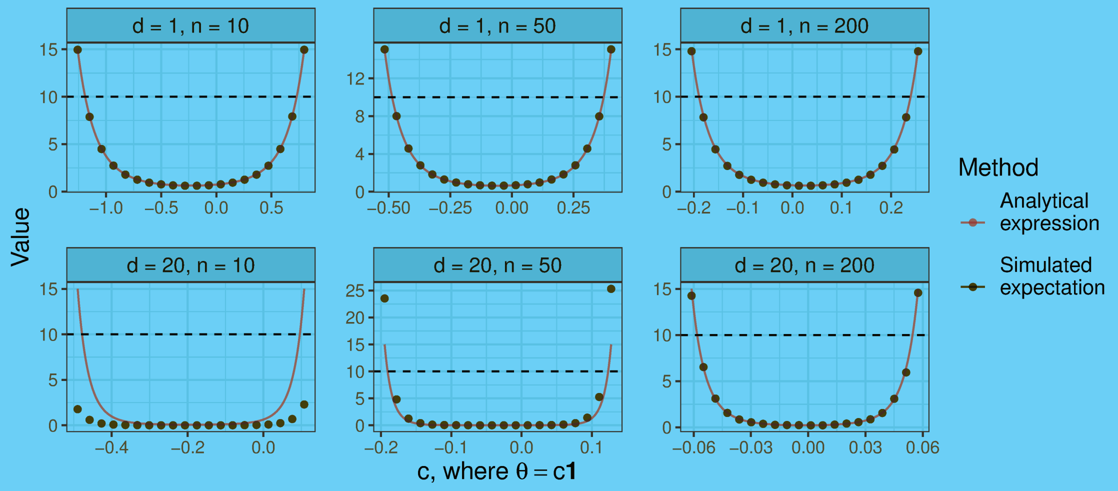

Fig. 2. Analytical (red curve) and simulated (black dots) approximations of the limiting test statistic $\lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} T_{n,b}(\theta)$ at various dimensions d and numbers of observations n. The test points equal $\theta = c\vec{1}$ for various c. The horizontal dashed black line at $1/0.1$ is the cutoff for an $\alpha = 0.1$ confidence region.

the subsampling confidence region is approximately

$$C_n^{\text{subsplit}}(\alpha) = \left\{\theta \in \Theta : \lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right) < \frac{1}{\alpha}\right\}$$

$$\approx \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < \frac{10}{3n}\log\left(\frac{(5/2)^{d/2}}{\alpha}\right)\right\}. \tag{5}$$

Figure 2 validates (4) as a reasonable approximation. We simulate one sample $Y_1, \ldots, Y_n \sim N(0, I_d)$ at $d = 1$ and $d = 20$, where $n = 1000$. We consider θ values of the form $\theta = c\vec{1}$. Through $B = 100,000$ subsampling simulations, we estimate

$$\mathbb{E}\{T_n(\theta) \mid \mathcal{D}\} \approx \frac{1}{B}\sum_{b=1}^{B} \exp\left(-\frac{n}{4}\|\overline{Y}_{0,b} - \overline{Y}_{1,b}\|^2 + \frac{n}{4}\|\overline{Y}_{0,b} - \theta\|^2\right).$$

The black dots represent this average at each value of c, and the red curve traces out $\exp((3n/10)\|\overline{Y} - \theta\|^2)(2/5)^{d/2}$ from (4). Except for the most difficult setting of $(d = 20, n = 10)$, the simulated and analytical estimates align well. At $\alpha = 0.1$, the confidence region includes all values of θ such that the test statistic is at most $1/0.1$. The horizontal dashed black line represents this value. Thus, test statistics constructed from the simulated and analytical approaches would produce similar confidence regions.

3. Comparison of universal LRT sets

3.1. *Optimal split proportions*

We have been assuming that the universal LRTs place $n/2$ observations in $\mathcal{D}_0$ and $n/2$ observations in $\mathcal{D}_1$. The statement $\mathbb{E}_{\theta^*}\{T_n(\theta^*)\} \leq 1$ holds regardless of the proportion of observations in $\mathcal{D}_0$ versus $\mathcal{D}_1$, though. Let p_0 denote the proportion of observations that we place in $\mathcal{D}_0$.

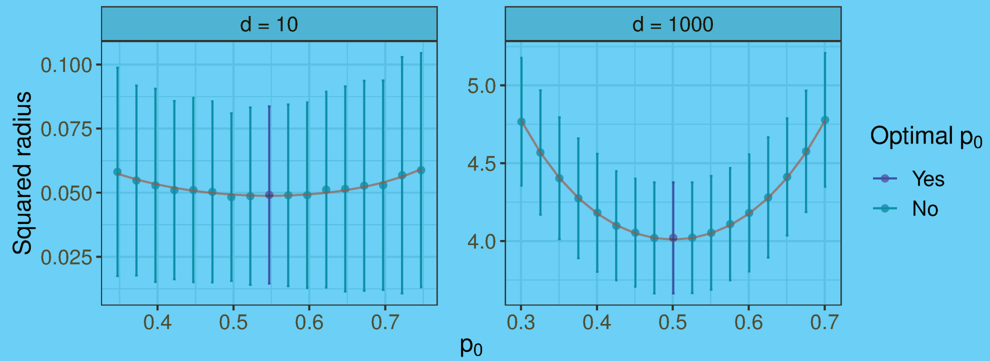

Fig. 3. Squared radius of multivariate normal split LRT with varying p_0. We simulate $Y_1, \ldots, Y_{1000} \sim N(0, I_d)$ and compute the split LRT region at varying p_0. We repeat this simulation 1000 times. At each p_0, the circular point is the mean squared radius and the error bar represents the mean squared radius $\pm$ 1.96 standard deviations. Blue points/lines correspond to p_0^*. The red curve is the expected squared radius. (See Theorem 3 proof in the supplement for a derivation of the expected squared radius at p_0.)

Theorem 3. *Let $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. The splitting proportion that minimizes $\mathbb{E}[r^2\{C_n^{split}(\alpha)\}]$ is*

$$p_0^* = 1 - \frac{\sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)} - 2d}{4\log\left(\frac{1}{\alpha}\right)}. \tag{6}$$

As $d \to \infty$ for fixed α, the optimal split proportion p_0^* converges to 0.5. See Section S1 of the supplement for a proof of Theorem 3 and a derivation of this fact. Alternatively, as $\alpha \to 0$ for fixed d, the proportion p_0^* converges to 1, suggesting that one should use nearly all data for likelihood estimation. This is not an issue for reasonable α levels, though. For instance, at $d = 1$, one would need to set $\alpha < \exp(-40)$ to produce an optimal split proportion p_0^* that exceeds 0.90.

Figure 3 shows the average squared radius of the split LRT at p_0^* and at surrounding choices of p_0. The expected squared radius (red curve) is more sensitive to changes in p_0 at higher values of d. That is, use of the optimal p_0^* has a greater effect on the split LRT squared radius in higher dimensions. In high dimensions, though, p_0^* is close to 0.5. It is thus a reasonable choice to use $p_0 = 0.5$ in all dimensions. We use $p_0 = 0.5$ for all remaining analyses.

In the cross-fit case, we conjecture that $p_0 = 0.5$ minimizes the expected squared diameter. Simulations in Section S3 of the supplement support this claim. Intuitively, since the cross-fit approach uses both $\mathcal{D}_0$ and $\mathcal{D}_1$ once for parameter estimation and once for likelihood computation, we should not gain any efficiency by using unbalanced sets.

3.2. *Split versus cross-fit volume*

In Fig. 1, we see that the cross-fit LRT set volume is less than the split LRT set volume, but $C_n^{\text{CF}}(\alpha)$ is not a subset of $C_n^{\text{split}}(\alpha)$. Nevertheless, it holds that Volume$\{C_n^{\text{CF}}(\alpha)\} \leq$ Volume$\{C_n^{\text{split}}(\alpha)\}$.

Theorem 4. *Suppose $Y_1, \ldots, Y_n$ are iid observations from $N(\theta^*, I_d)$. Split the sample such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $\frac{n}{2}$ observations. Use $\mathcal{D}_0$ and $\mathcal{D}_1$ to define the split and cross-fit sets. Then Volume$\{C_n^{CF}(\alpha)\} \leq$ Volume$\{C_n^{split}(\alpha)\}$. Equality holds only when $\overline{Y}_0 = \overline{Y}_1$.*

Briefly, the proof of Theorem 4 constructs a spherical region centered at $\overline{Y}$ with radius equal to the split LRT radius. The cross-fit set is a subset of this re-centered split LRT region, so the volume of the cross-fit LRT set is bounded above by the volume of the split LRT set. If $\overline{Y}_0 = \overline{Y}_1$, then the split and cross-fit

LRT sets are equivalent and have equal volume. The fact that equal volume holds only when $\overline{Y}_0 = \overline{Y}_1$ relies on the strict convexity of the squared L_2 norm and the exponential function. See Section S1 of the supplement for a complete proof.

Theorem 4 proves that the cross-fit LRT approach improves over the split LRT by constructing provably smaller confidence regions. Out of all universal methods, our simulations have shown that the subsampling approach tends to produce the smallest sets. Constructing a subsampling region can be computationally intensive, though, especially when the limiting subsampling test statistic is intractable. The cross-fit approach may be a reasonable compromise in settings where repeated subsampling is computationally prohibitive.

3.3. *Comparative size in high dimensions*

Figure 1 demonstrated the appearance of the four LRT regions in the $d = 2$ case at $\alpha = 0.1$. We observe that the classical LRT and the split LRT produce the smallest and largest confidence regions, respectively. While the split LRT region's radius appears to be approximately twice the classical LRT region's radius, we consider whether the ratio of their squared radii diverges in high dimensions or for very small α. We characterize the ratio of squared radii in terms of the expected ratio. The expected squared radius of $C_n^{\text{split}}(\alpha)$ is

$$\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}] = (4/n)\log(1/\alpha) + (4/n)d. \tag{7}$$

Thus, the expected ratio of the split LRT squared radius over the classical LRT radius is

$$\frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} = \frac{(4/n)\log(1/\alpha) + (4/n)d}{c_{\alpha,d}/n} = \frac{4\log(1/\alpha) + 4d}{c_{\alpha,d}}. \tag{8}$$

For $d \geq 2$ and $\alpha \leq 0.17$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d - \frac{5}{2}}. \tag{9}$$

For $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$,

$$\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \frac{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]}{r^2\{C_n^{\text{LRT}}(\alpha)\}} \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}}. \tag{10}$$

See Section S2 of the supplement for derivations of (7), (9), and (10). The derivation of (9) relies on chi square quantile bounds from Theorem A and Proposition 5.1 of Inglot (2010). The derivation of the upper bound in (10) involves a bound from Section 2.1 of Polland (2015). The restrictions on α and d are necessary for the upper bounds to be valid. The lower bound is valid for any $d \geq 1$ and $\alpha \in (0, 1)$. The upper and lower bounds both converge to 4 as $d \to \infty$. In addition, all bounds converge to 2 as $\alpha \to 0$. Figure 4 shows the true value of $\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}] \,/\, r^2\{C_n^{\text{LRT}}(\alpha)\}$ as well as the proved lower and upper bounds on this expectation at $d = 10$ and $d = 100,000$. We observe that the bounds converge to 2 for very small α relative to the dimension, and we observe that the bounds converge to 4 for high dimensions relative to α. Interestingly, we see that the expected value of the ratio is not monotone increasing in α.

Furthermore, this ratio of squared radii is less than 4 with probability approximately $1 - \alpha$ in high dimensions. Theorem 5 formalizes this result. See Section S1 of the supplement for a proof.

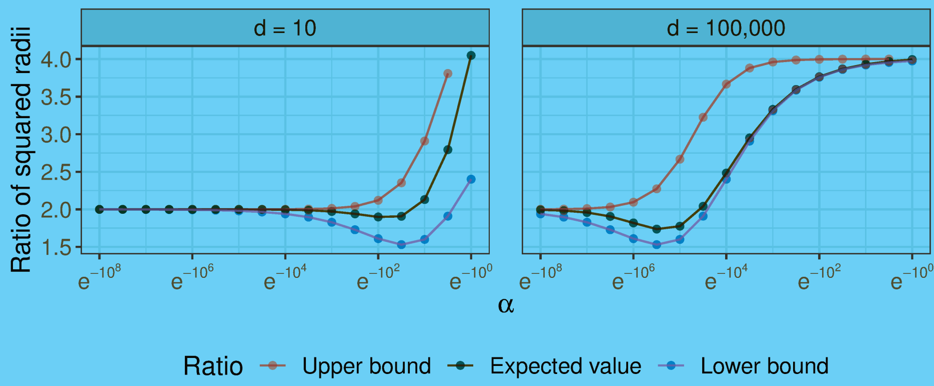

Fig. 4. Expectation (black), lower bound (blue), and upper bound (red) of $E\left[r^2\left\{C_n^{\text{split}}(\alpha)\right\}/r^2\{C_n^{\text{LRT}}(\alpha)\}\right]$. The expected value equals the expression from (8). The lower and upper bounds correspond to the bounds in (9). Data points correspond to values at $\alpha = \exp(-10^x)$ for x from 8 to 0 in increments of -0.5.

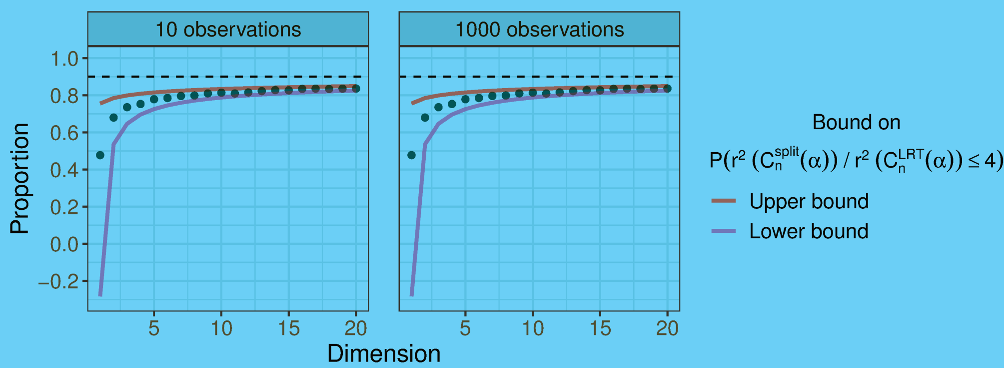

Fig. 5. We perform 10,000 simulations in which we simulate a data sample $Y_1, \ldots, Y_{1000} \sim N(0, I_2)$, construct the split and classical LRT confidence sets, and compute the squared radii. The points represent the proportion of these simulations in which $r^2\{C_n^{\text{split}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\} \leq 4$. The red and blue curves are the lower and upper bounds on $\mathbb{P}[r^2\{C_n^{\text{split}}(\alpha)\} \,/\, r^2\{C_n^{\text{LRT}}(\alpha)\} \leq 4]$ from Theorem 5 at $\alpha = 0.1$.

Theorem 5. *Assume $c_{\alpha,d} + \log(\alpha) > d - 2$. Let $f_d(x)$ be the probability density function of the χ^2_d distribution, and let $c_{\alpha,d}$ be the upper α quantile of the χ^2_d distribution. Then*

$$\mathbb{P}\left[r^2\{C_n^{\mathit{split}}(\alpha)\}/r^2\{C_n^{\mathit{LRT}}(\alpha)\} \leq 4\right] \geq 1 - \alpha - \log(1/\alpha) f_d\{c_{\alpha,d} + \log(\alpha)\}$$
$$\textit{and} \quad \mathbb{P}\left[r^2\{C_n^{\mathit{split}}(\alpha)\}/r^2\{C_n^{\mathit{LRT}}(\alpha)\} \leq 4\right] \leq 1 - \alpha - \log(1/\alpha) f_d(c_{\alpha,d}).$$

Figure 5 explores the bounds from Theorem 5. We see that the result from Theorem 5 is more informative in higher dimensions, where the upper and lower bounds are closer to each other. Both theoretically and empirically, the ratio of squared radii $r^2\{C_n^{\text{split}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\}$ is less than 4 with probably slightly below $1 - \alpha$ in higher dimensions.

From (5) and (7), we can see that

$$\frac{r^2\{C_n^{\text{subsplit}}(\alpha)\}}{\mathbb{E}[r^2\{C_n^{\text{split}}(\alpha)\}]} \approx \frac{5}{6}\left\{\frac{(d/2)\log(5/2) + \log(1/\alpha)}{d + \log(1/\alpha)}\right\}. \tag{11}$$

Combining (9) and (11), $r^2\{C_n^{\text{subsplit}}(\alpha)\}/r^2\{C_n^{\text{LRT}}(\alpha)\}$ is approximately $4(5/12)\log(5/2) \approx 3/2$ as $d \to \infty$, and the ratio is approximately $2(5/6) = 5/3$ as $\alpha \to 0$. Recall that the classical LRT cutoff is dimension dependent and uses the exact distribution's quantile, while the universal LRT cutoff is dimension independent. Regardless, in the extreme cases of $d \to \infty$ or $\alpha \to 0$, the ratio of the classical LRT region's radius to the subsampling universal LRT region's radius is less than 2.

3.4. *Power*

While the universal methods provide conservative confidence regions for θ^*, we establish that the universal tests can still have high power. Suppose we wish to test $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ at level $1 - \alpha$. We reject H_0 if $0 \notin C_n(\alpha)$, where $C_n(\alpha)$ is the confidence set defined by some likelihood ratio test. The power of the test at $\theta^* \neq 0$ is $\mathbb{P}_{\theta^*}\{0 \notin C_n(\alpha)\}$.

First, we consider the classical LRT, stated in (1). The power of the classical LRT at θ^* is

$$\text{Power}\{C_n^{\text{LRT}}(\alpha); \theta^*\} = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > c_{\alpha,d}/n\right) \approx \Phi\left\{\frac{d + n\|\theta^*\|^2 - c_{\alpha,d}}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right\}. \tag{12}$$

We can find a similar representation for the approximate power of the limiting subsampling LRT as $B \to \infty$:

$$\begin{aligned}\text{Power}\{C_n^{\text{subsplit}}(\alpha); \theta^*\} &\approx \mathbb{P}_{\theta^*}\left[n\|\overline{Y}\|^2 \geq \frac{10}{3}\log\left\{\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right\}\right] \\ &\approx \Phi\left(\frac{1}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\left[d + n\|\theta^*\|^2 - \frac{10}{3}\log\left\{\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right\}\right]\right). \end{aligned} \tag{13}$$

Since $n\|\overline{Y}\|^2 \sim \chi^2\left(df = d, \lambda = n\|\theta^*\|^2\right)$, (12) and (13) use the normal approximation to the non-central χ^2 distribution with a large noncentrality parameter λ (Chun & Shapiro, 2009). See Section S2 of the supplement for derivations of (12) and (13).

The power of the split LRT is

$$\text{Power}\{C_n^{\text{split}}(\alpha); \theta^*\} = \mathbb{P}_{\theta^*}\left\{\|\overline{Y}_0\|^2 \geq (4/n)\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$$

and the power of the cross-fit LRT is

$$\text{Power}\{C_n^{\text{CF}}(\alpha); \theta^*\} = \mathbb{P}_{\theta^*}\left[\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2\right)\left\{\exp\left(\frac{n}{4}\|\overline{Y}_0\|^2\right) + \exp\left(\frac{n}{4}\|\overline{Y}_1\|^2\right)\right\} \geq \frac{2}{\alpha}\right].$$

As $n\|\theta^*\|^2 \to \infty$ for fixed α, the power of the tests approaches 1. Importantly, this shows that although the universal methods are conservative, they will all have high power for sufficiently large n or for $\|\theta^*\|$ sufficiently far from 0. As $\alpha \to 0$, the power approaches 0.

Figure 6 plots the power of the LRTs against $\|\theta^*\|^2$. (Each vector θ^* has the form $c\vec{1}$.) This figure uses the standard normal CDF approximation to the non-central χ^2 CDF to plot the classical and subsampling LRT power. We use simulations to approximate the power of the split and cross-fit LRTs. For a given value of θ^*, we simulate $n = 1000$ observations $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. We construct split LRT and cross-fit LRT confidence sets from this sample. Then we test whether $\theta = 0$ is in each confidence set. We repeat this procedure 5000 times at each θ^*, and each procedure's estimated power at θ^* is the proportion of times that $0 \notin C_n(\alpha)$.

As we would expect, the power is higher when θ^* is farther from 0. In addition, the classical LRT has the highest power, followed in order by the subsampling LRT, the cross-fit LRT, and the split LRT. Interestingly, at $d = 1$ the subsampling and cross-fit LRT have nearly identical (approximate) power. As d increases, the difference between the subsampling and cross-fit LRT power increases.

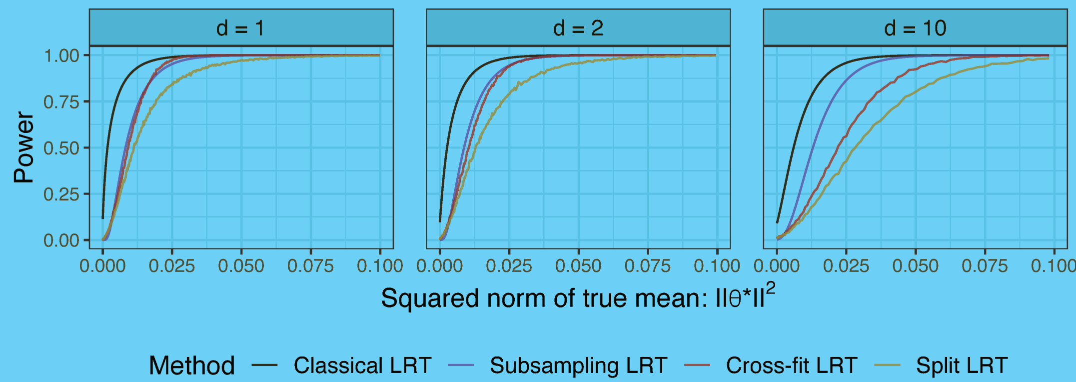

Fig. 6. Estimated power of classical LRT (black), limiting subsampling LRT (blue), cross-fit LRT (red), and split LRT (orange). We are testing $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$ across varying true $\|\theta^*\|^2$. We use the standard normal CDF approximation for the classical and subsampling LRT power calculations, and we use simulations to estimate the cross-fit and split LRT power.

4. Example: hypothesis testing a doughnut null set

Instead of presenting a simulation which further confirms our theoretical findings, we instead present here an example of a nontrivial testing problem that appears to be beyond the current reach of our mathematical analysis. Below, a procedure based on universal inference can have higher power than a more standard intersection approach using the classical, exact confidence set, motivating the need for further study of the pros and cons of such methods.

Suppose we observe an iid sample $Y_1, \dots, Y_n \sim N(\theta^*, I_d)$, and we wish to test

$$H_0 : \|\theta^*\| \in [0.5, 1.0] \text{ versus } H_1 : \|\theta^*\| \notin [0.5, 1.0].$$

Then $\Theta_0 = \{\theta \in \mathbb{R}^d : \|\theta\| \in [0.5, 1.0]\}$ and $\Theta_1 = \{\theta \in \mathbb{R}^d : \|\theta\| \notin [0.5, 1.0]\}$. The nonconvex structure of Θ_0 makes it unclear how to construct a valid test based on a limiting distribution. Nevertheless, we can use alternative methods, including universal inference tools, to construct valid hypothesis tests for $H_0 : \|\theta^*\| \in [0.5, 1.0]$. We compare three approaches to this test.

Approach 1: Intersect confidence set with Θ_0. $C_n^{\mathrm{LRT}}(\alpha) = \{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \leq c_{\alpha,d}/n\}$ is a level $1 - \alpha$ confidence set for θ^*, where $c_{\alpha,d}$ is the upper α quantile of the χ^2_d distribution. Suppose we reject H_0 if and only if $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset$. We can see that this test has valid type I error control. Assume $\theta^* \in \Theta_0$. Then

$$\begin{aligned}\mathbb{P}_{\theta^*}\left\{C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset\right\} &\leq \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\mathrm{LRT}}(\alpha) \cup \theta^* \notin \Theta_0\right\} \\ &= \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\mathrm{LRT}}(\alpha)\right\} \\ &= \alpha.\end{aligned}$$

To implement this test, we need to check whether the intersection $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0$ is empty. First, we set $\widehat{\theta}^{\mathrm{proj}}$ to the projection of $\overline{Y}$ onto Θ_0. That is,

$$\widehat{\theta}^{\mathrm{proj}} = \begin{cases} 0.5\,\overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| < 0.5 \\ \overline{Y} & : \|\overline{Y}\| \in [0.5, 1.0] \\ \overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| > 1. \end{cases}$$

Now $\widehat{\theta}^{\mathrm{proj}}$ minimizes $\|\theta - \overline{Y}\|^2$ out of all $\theta \in \Theta_0$. So $C_n^{\mathrm{LRT}}(\alpha) \cap \Theta_0 = \emptyset$ if and only if $\widehat{\theta}^{\mathrm{proj}} \notin C_n^{\mathrm{LRT}}(\alpha)$.

Approach 2: Subsampled split LRT. To implement the subsampled split LRT, we repeatedly split the observations into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$. Let $\widehat{\theta}_{1,b}$ be any parameter estimated on the data in $\mathcal{D}_{1,b}$. Let $\widehat{\theta}_{0,b}^{\text{split}}$ be the MLE under $H_0 : \|\theta^*\| \in [0.5, 1.0]$, estimated on the data in $\mathcal{D}_{0,b}$. Table 1 presents the chosen expression for $\widehat{\theta}_{1,b}$ and the MLE expression for $\widehat{\theta}_{0,b}^{\text{split}}$. The subsampled split LRT rejects H_0 if $B^{-1}\sum_{b=1}^{B} U_{n,b} \geq 1/\alpha$, where

$$U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \,/\, \mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{split}}) = \prod_{Y_i \in \mathcal{D}_{0,b}} \{p_{\widehat{\theta}_{1,b}}(Y_i) \,/\, p_{\widehat{\theta}_{0,b}^{\text{split}}}(Y_i)\}.$$

Approach 3: Subsampled hybrid LRT. As an alternative to the split LRT, Wasserman et al. (2020) establish a test based on the reversed information projection (RIPR); also see Grünwald et al. (2020). We first define the RIPR, following Definition 4.2 of the PhD thesis by Li (1999). Let Q be a distribution with density q, and let $\mathcal{P}_\Theta$ be a convex set of densities (or redefine it as its convex hull). Let $D_{\text{KL}}(\cdot \,\|\, \cdot)$ be the Kullback-Leibler divergence. The RIPR of q onto $\mathcal{P}_\Theta$ is a (sub-)density p^* such that for arbitrary sequences p_n in $\mathcal{P}_\Theta$, $D_{\text{KL}}(q \,\|\, p_n) \to \inf_{\theta\in\Theta} D_{\text{KL}}(q \,\|\, p_\theta)$ implies $\log(p_n) \to \log(p^*)$ in $L^1(Q)$. Lemma 4.1 of Li (1999) proves that p^* exists and is unique; further, p^* satisfies $D_{\text{KL}}(q \,\|\, p^*) = \inf_{\theta\in\Theta} D_{\text{KL}}(q \,\|\, p_\theta)$, and if $Y \sim q$, then for all $\theta \in \Theta$, $\mathbb{E}_q\{p_\theta(Y)/p^*(Y)\} \leq 1$.

Using similar logic to Theorem 1, Wasserman et al. (2020) apply this property to construct a split RIPR LRT. Let $\mathcal{P}_{\Theta_0}$ be the set of all densities in H_0 (or its convex hull). Suppose $\widehat{\theta}_1$ is an estimator constructed on $\mathcal{D}_1$. Let p_0^* be the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$. Note that if the true $p_{\theta^*} \in \mathcal{P}_{\Theta_0}$, then $\mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_0^*(Y)\} = \mathbb{E}_{\widehat{\theta}_1}\{p_{\theta^*}(Y)/p_0^*(Y)\} \leq 1$. Then a level α hypothesis test rejects H_0 if $R_n \geq 1/\alpha$, where

$$R_n = \prod_{Y_i \in \mathcal{D}_0} \{p_{\widehat{\theta}_1}(Y_i) \,/\, p_0^*(Y_i)\}.$$

This test is valid because if $\theta^* \in \Theta_0$, then $\mathbb{P}_{\theta^*}(R_n \geq 1/\alpha) \leq \alpha\mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_0^*(Y)\} \leq \alpha$. Furthermore, note that the RIPR test statistic will always exceed the split LRT statistic when the two tests use the same numerator, since the split LRT denominator maximizes the likelihood under H_0 on $\mathcal{D}_0$. Thus, the RIPR test will have higher power than the split LRT. (More generally, one can project $p_{\widehat{\theta}_1}^{|\mathcal{D}_0|}$ onto $\mathcal{P}_{\Theta_0}^{|\mathcal{D}_0|}$, but we omit this discussion for brevity.)

In the doughnut test setting, we let $\mathcal{P}_{\Theta_0}$ be the set of all convex combinations of $N(\theta, I_d)$ densities such that $\|\theta\| \in [0.5, 1]$. To implement the subsampled hybrid LRT for this test, we also repeatedly split the observations into $\mathcal{D}_{0,b}$ and $\mathcal{D}_{1,b}$. Depending on the value of $\|\overline{Y}_{1,b}\|$, we take one of three approaches:

1. If $\|\overline{Y}_{1,b}\| < 0.5$, use the split LRT on the b^{th} subsample. We define $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}^{\text{split}}$ as in Table 1, and the split LRT statistic is $U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{split}})$.
2. If $\|\overline{Y}_{1,b}\| \in [0.5, 1]$, set the b^{th} subsample's test statistic to 1.
3. If $\|\overline{Y}_{1,b}\| > 1$, use the RIPR LRT on the b^{th} subsample. We define $\widehat{\theta}_{1,b}$ and $\widehat{\theta}_{0,b}^{\text{RIPR}}$ as in Table 1, and the RIPR statistic is $R_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b})/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{RIPR}})$.

Theorem 6 defines a valid test based on this approach.

Theorem 6. *In the doughnut null hypothesis test setting, assume the subsampled test statistics $U_{n,b}$ and $R_{n,b}$, $1 \leq b \leq B$, as defined above. A valid level α test rejects H_0 when*

$$\frac{1}{B}\sum_{b=1}^{B}\Big\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\Big\} \geq 1/\alpha.$$

Table 1. *Requirements and choices for the numerator and denominator in a single subsample of the split LRT and RIPR LRT statistics*

Method	Split LRT	RIPR LRT
Restrictions on use	None	$\lVert \overline{Y}_1 \rVert > 1$. (Computational restriction. RIPR unknown for $\lVert \overline{Y}_1 \rVert = \lVert \widehat{\theta}_1 \rVert < 0.5$.)
Numerator	$p_{\widehat{\theta}_1}$, where $\widehat{\theta}_1$ is any parameter fit on $\mathcal{D}_1$.	$p_{\widehat{\theta}_1}$, where $\widehat{\theta}_1$ is any parameter fit on $\mathcal{D}_1$.
Fitted value	Choose $\widehat{\theta}_1 = \overline{Y}_1$.	Choose $\widehat{\theta}_1 = \overline{Y}_1$.
Denominator	$p_{\widehat{\theta}_0}$, where $\widehat{\theta}_0$ is the MLE under H_0, constructed from $\mathcal{D}_0$.	p_0^* is the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$.
Fitted value	No choices. $\widehat{\theta}_0^{\text{split}} = \begin{cases} 0.5\left(\overline{Y}_0/\lVert \overline{Y}_0 \rVert\right) & : \lVert \overline{Y}_0 \rVert < 0.5 \\ \overline{Y}_0 & : \lVert \overline{Y}_0 \rVert \in [0.5, 1] \\ \overline{Y}_0/\lVert \overline{Y}_0 \rVert & : \lVert \overline{Y}_0 \rVert > 1 \end{cases}$	No choices. Since $\lVert \widehat{\theta}_1 \rVert > 1$, $p_0^* = p_\theta$, where $\theta = \widehat{\theta}_0^{\text{RIPR}} = \widehat{\theta}_1/\lVert \widehat{\theta}_1 \rVert$.

To justify the hybrid approach, recall that the RIPR test will have higher power than the split LRT when it is possible to implement the RIPR. Based on the construction of $\widehat{\theta}_{1,b}$, if $\lVert \overline{Y}_{1,b} \rVert > 1$, then $\lVert \widehat{\theta}_{1,b} \rVert > 1$. In this setting, the proof of Theorem 6 shows that the density p_θ, with $\theta = \widehat{\theta}_{1,b}/\lVert \widehat{\theta}_{1,b} \rVert$, is the RIPR of $\widehat{\theta}_{1,b}$ onto $\mathcal{P}_{\Theta_0}$. On the other hand, it is unclear how to implement the RIPR when $\lVert \overline{Y}_{1,b} \rVert < 0.5$, in which case $\lVert \widehat{\theta}_{1,b} \rVert < 0.5$. The hybrid approach allows us to use the RIPR when it is implementable, and it relies on the split LRT to provide a valid test when the RIPR is not implementable.

Figure 7 shows the simulated power of these three tests of $H_0 : \lVert \theta^* \rVert \in [0.5, 1.0]$ versus $H_1 : \lVert \theta^* \rVert \notin [0.5, 1.0]$. The intersection method and the subsampled hybrid LRT have the highest power. Interestingly, out of those two methods, the test with higher power varies across dimensions. When $d = 2$ or $d = 1000$, the simulated power of the subsampled hybrid LRT is less than (or equal to) the power of the standard intersection approach. At the intermediate dimensions of $d = 10$ and $d = 100$, the simulated power of the subsampled hybrid LRT is greater than (or equal to) the power of the standard intersection approach. The latter two cases show that even in the Gaussian setting, hypothesis tests based on a universal LRT can have higher power than tests based on the exact confidence set. When $\lVert \theta^* \rVert < 0.5$, the hybrid test and the split test have approximately the same power. When $\lVert \theta^* \rVert > 1$, the hybrid test has higher power than the split test. We see that the intersection method always has higher power than the subsampled split LRT. One might consider whether we could combine the RIPR with the intersection method instead of combining the RIPR with the split LRT. It is unclear, though, how to construct a valid test from one approach that uses sample splitting and subsampling (RIPR) and a second approach that uses neither (intersection).

We can provide a partial theoretical justification for Fig. 7. For one, it is possible to derive an exact formula for the power of the intersection approach. Using the fact that $n\lVert \overline{Y} \rVert^2$ follows a non-central χ^2 distribution, we can write the power of the intersection method in terms of the non-central χ^2 CDF. When $d = 100$ or $d = 1000$, the hybrid method has no power at $\lVert \theta^* \rVert = 0$, though we would expect this case to have the highest power out of $\lVert \theta^* \rVert < 0.5$. At $d = 100$ and $\lVert \theta^* \rVert = 0$, the hybrid method satisfies $\lVert \overline{Y}_{1,b} \rVert < 0.5$ in most simulations, but the test statistic is too small to reject H_0. At $d = 1000$ and $\lVert \theta^* \rVert = 0$, $(n/2)\lVert \overline{Y}_{1,b} \rVert^2 \sim \chi^2_d$ is approximately d (Dasgupta & Schulman, 2007, Lemma 2). Hence $\lVert \overline{Y}_{1,b} \rVert \approx \sqrt{2}$, which means the hybrid approach selects the "incorrect" case of $\lVert \overline{Y}_{1,b} \rVert > 1$. This test also has approximately zero power. See Section S4 of the supplement for more details. In addition, for any given subsample, the hybrid LRT power is provably greater than or equal to the split LRT power. This holds because the RIPR test statistic is always larger than the split test statistic when both tests use the

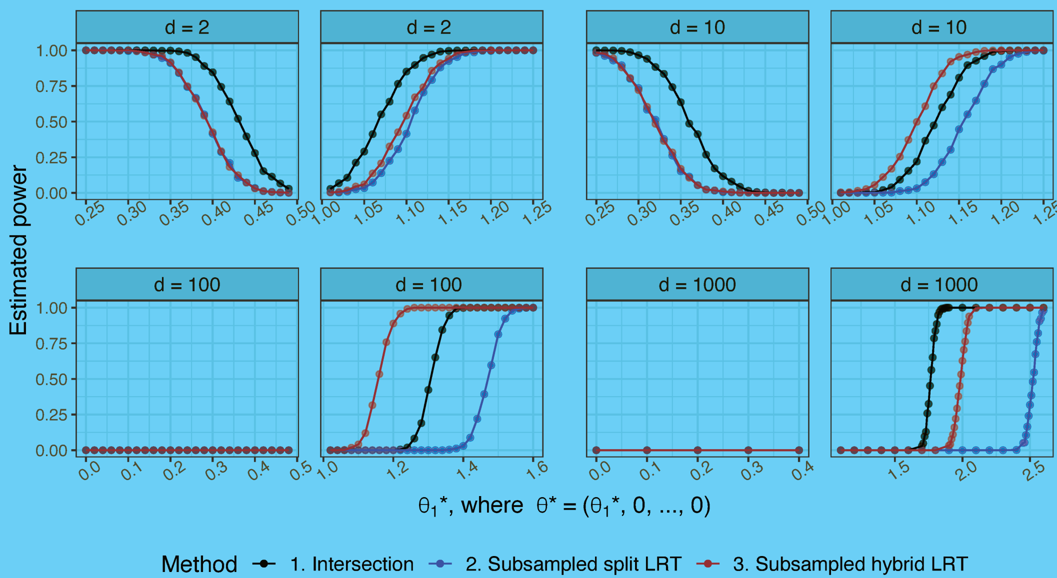

Fig. 7. Estimated power of $H_0 : \|\theta^*\| \in [0.5, 1.0]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1.0]$ using the intersection (black), subsampled split LRT (blue), and subsampled hybrid LRT (red) methods. In these simulations, we set $\theta^* = (\theta_1^*, 0, \ldots, 0)$. The x-axis is the value of $\theta_1^* = \|\theta^*\|$ for each simulation. For each dimension, the left panel satisfies $\|\theta^*\| < 0.5$, and the right panel satisfies $\|\theta^*\| > 1$. We set $\alpha = 0.10$ and $n = 1000$, and we perform 1000 simulations at each value of $\|\theta^*\|$. We subsample $B = 100$ times.

same numerator (Wasserman et al., 2020). The theoretical justification behind the relative power of the intersection and subsampled hybrid methods remains an open question, since the power of the latter method is not easily tractable.

5. Discussion

The recent development of the universal LRT provides a hypothesis testing framework that is valid in finite samples and does not rely on regularity conditions. We have explored the performance of several universal LRT variants in the simple but fundamental case of testing for the mean θ^* when data arise from a $N(\theta^*, I_d)$ distribution. We have seen that even in high dimensions or for very small α, the ratio of the radius of the limiting subsampling universal LRT confidence set over an exact confidence set is less than 2. While the universal method tests the likelihood ratio against a dimension-independent cutoff, the universal LRT can still exhibit reasonable performance in high dimensions.

Future research directions may focus on settings where hypothesis tests were previously intractable or only asymptotically valid. Researchers can apply the universal LRT in any setting where it is possible to write a likelihood ratio or, more generally, upper bound the maximum likelihood under the null hypothesis. This allows for the development of valid tests for the number of components in mixture models and for log-concavity of the underlying density. Additionally, we have shown proof of concept that the universal LRT can be more powerful than existing valid tests. In the Gaussian setting, this phenomenon may apply more generally across other tests of non-convex null parameter spaces. Wasserman et al. (2020) also describe how the universal LRT can be used to test independence versus conditional independence in a Gaussian setting. Recent work by Guo & Richardson (2020) also provides a valid test in that setting, but the relative power of these two approaches is currently unknown.

Acknowledgement

RD's research is supported by the National Science Foundation Graduate Research Fellowship Program under Grant Nos. DGE 1252522 and DGE 1745016. AR's research is supported by the Adobe Faculty Research Award, an ARL Large Grant, and the National Science Foundation under Grant Nos. DMS 2053804, DMS 1916320, and DMS (CAREER) 1945266. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. This work used the Extreme Science and Engineering Discovery Environment (XSEDE) (Towns et al., 2014), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system (Nystrom et al., 2015), which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC). This work made extensive use of the `R` statistical software (R Core Team, 2021), as well as the `data.table` (Dowle & Srinivasan, 2021), `ggConvexHull` (Martin, 2017), `MASS` (Venables & Ripley, 2002), `Rcpp` (Eddelbuettel & François, 2011; Eddelbuettel, 2013; Eddelbuettel & Balamuta, 2018), and `tidyverse` (Wickham et al., 2019) packages.

References

Axelrod, B., Diakonikolas, I., Stewart, A., Sidiropoulos, A. & Valiant, G. (2019). A polynomial time algorithm for log-concave maximum likelihood via locally exponential families. In *Advances in Neural Information Processing Systems*.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences* **160**, 268–282.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience* **14**, 365–376.

Chen, J., Li, P. et al. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics* **37**, 2523–2542.

Chun, S. Y. & Shapiro, A. (2009). Normal versus noncentral chi-square asymptotics of misspecified models. *Multivariate Behavioral Research* **44**, 803–827.

Cule, M., Samworth, R. & Stewart, M. (2010). Maximum likelihood estimation of a multi-dimensional log-concave density. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **72**, 545–607.

Dasgupta, S. & Schulman, L. J. (2007). A probabilistic analysis of EM for mixtures of separated, spherical Gaussians. *Journal of Machine Learning Research* **8**, 203–226.

Dowle, M. & Srinivasan, A. (2021). *data.table: Extension of 'data.frame'*. R package version 1.14.0.

Eddelbuettel, D. (2013). *Seamless R and C++ Integration with Rcpp*. New York: Springer. ISBN 978-1-4614-6867-7.

Eddelbuettel, D. & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician* **72**, 28–36.

Eddelbuettel, D. & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software* **40**, 1–18.

Grünwald, P., de Heide, R. & Koolen, W. M. (2020). Safe testing. In *2020 Information Theory and Applications Workshop (ITA)*. IEEE.

Guo, F. R. & Richardson, T. S. (2020). On testing marginal versus conditional independence. *arXiv preprint arXiv:1906.01850v2* .

Hájek, J. (1960). Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences* **5**, 361–374.

Hartigan, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Barkeley Conference in Honor of J. Neyman and J. Kiefer*, vol. 2.

Ildstad, S. T., Evans Jr, C. H. et al. (2001). *Small clinical trials: Issues and challenges*. National Academies Press.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics* **30**, 339–351.

Lehmann, E. L. (2012). On likelihood ratio tests. In *Selected Works of EL Lehmann*. Springer, pp. 209–216.

Li, P. & Chen, J. (2010). Testing the order of a finite mixture. *Journal of the American Statistical Association* **105**, 1084–1092.

Li, Q. (1999). *Estimation of mixture models*. Yale University. http://www.stat.yale.edu/~arb4/students_files/JonathanLiThesis.pdf.

Li, X. & Ding, P. (2017). General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association* **112**, 1759–1769.

Martin, C. A. (2017). *ggConvexHull: Add a convex hull geom to ggplot2*. R package version 0.1.0.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **36**, 318–324.

McMahon, A. W., Watt, K., Wang, J., Green, D., Tiwari, R. & Burckart, G. J. (2016). Stratification, hypothesis testing, and clinical trial simulation in pediatric drug development. *Therapeutic innovation & regulatory science* **50**, 817–822.

Medeiros, F. M. & Ferrari, S. L. (2017). Small-sample testing inference in symmetric and log-symmetric linear regression models. *Statistica Neerlandica* **71**, 200–224.

Nystrom, N. A., Levine, M. J., Roskies, R. Z. & Scott, J. R. (2015). Bridges: A Uniquely Flexible HPC Resource for New Communities and Data Analytics. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE '15. New York, NY, USA: Association for Computing Machinery.

Polland, D. (2015). A few good inequalities. In *Mini-empirical (draft)*.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R. & Wilkins-Diehr, N. (2014). XSEDE: Accelerating Scientific Discovery. *Computing in Science & Engineering* **16**, 62–74.

van der Vaart, A. W. (2000). *Asymptotic statistics*, vol. 3. Cambridge university press.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S*. New York: Springer, 4th ed. ISBN 0-387-95457-0.

Wasserman, L., Ramdas, A. & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117**, 16880–16890.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K. & Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software* **4**, 1686.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62.

S1. PROOFS OF THEOREMS

Theorem 1. *$C_n^{split}(\alpha)$ is a valid $100(1-\alpha)\%$ confidence set for θ^*. As a consequence (and equivalently), when testing an arbitrary composite null $H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta \setminus \Theta_0$, rejecting H_0 when $\Theta_0 \cap C_n^{split}(\alpha) = \emptyset$ provides a valid level α hypothesis test. The latter rule reduces to rejecting if $T_n(\widehat{\theta}_0) \geq 1/\alpha$, where $\widehat{\theta}_0 \in \arg\max_{\theta\in\Theta_0} \mathcal{L}_0(\theta)$ is the null MLE.*

Proof. This result is due to Wasserman et al. (2020). To prove this fact, we show that $E_{\theta^*}[T_n(\theta^*) \mid \mathcal{D}_1] \leq 1$. First, we use only the data in $\mathcal{D}_1$ to fit a parameter $\widehat{\theta}_1$. Let $\mathcal{M}(\theta) = support(P_\theta)$ in $|\mathcal{D}_0|$-dimensional space. We see

$$
\begin{aligned}
\mathbb{E}_{\theta^*}[T_n(\theta^*) \mid \mathcal{D}_1] &= \mathbb{E}_{\theta^*}\left[\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta^*)} \,\middle|\, \mathcal{D}_1\right] = \int_{\mathcal{M}(\theta^*)} \frac{\prod_{y_i\in\mathcal{D}_0} p_{\widehat{\theta}_1}(y_i)}{\prod_{y_i\in\mathcal{D}_0} p_{\theta^*}(y_i)} \prod_{y_i\in\mathcal{D}_0} p_{\theta^*}(y_i) dy_i \\
&= \int_{\mathcal{M}(\theta^*)} \prod_{y_i\in\mathcal{D}_0} p_{\widehat{\theta}_1}(y_i) dy_i \leq \int_{\mathcal{M}(\widehat{\theta}_1)} \prod_{y_i\in\mathcal{D}_0} p_{\widehat{\theta}_1}(y_i) dy_i \\
&\stackrel{iid}{=} \int_{\mathcal{M}(\widehat{\theta}_1)} p_{\widehat{\theta}_1}\left(y_1, \ldots, y_{|\mathcal{D}_0|}\right) \prod_{y_i\in\mathcal{D}_0} dy_i = 1.
\end{aligned}
$$

Applying Markov's inequality and the above fact,

$$
\mathbb{P}_{\theta^*}\left(\theta^* \notin C_n^{\text{split}}(\alpha)\right) = \mathbb{P}_{\theta^*}\left(T_n(\theta^*) \geq 1/\alpha\right) \leq \alpha \mathbb{E}_{\theta^*}[T_n(\theta^*)] = \alpha \mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left[T_n(\theta^*) \mid \mathcal{D}_1\right]\right] \leq \alpha.
$$

This shows that $\theta^* \in C_n^{\text{split}}(\alpha)$ with probability at least $1-\alpha$. Alternatively, suppose we want to test $H_0 : \theta^* \in \Theta_0$ versus $H_1 : \theta^* \in \Theta \setminus \Theta_0$. We see that rejecting H_0 when $\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset$ provides a valid level α hypothesis test. Under H_0,

$$
\mathbb{P}_{\theta^*}\left\{\Theta_0 \cap C_n^{\text{split}}(\alpha) = \emptyset\right\} \leq \mathbb{P}_{\theta^*}\left\{\theta^* \notin \Theta_0 \cap C_n^{\text{split}}(\alpha)\right\} = \mathbb{P}_{\theta^*}\left\{\theta^* \notin C_n^{\text{split}}(\alpha)\right\} \leq \alpha.
$$

□

Before proving Theorem 2, we establish Lemma 1 and Lemma 2. We draw heavily on finite population central limit theorem results from Hájek (1960) and Li & Ding (2017). Lemma 1 combines key results from these two papers and adapts them to our setting.

Lemma 1. *Let $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$ be a sequence of datasets, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$. Define $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_{ni}$ and $\overline{Y}_{0,n} = \frac{2}{n}\sum_{Y_{ni}\in\mathcal{D}_{0,n}} Y_{ni}$. As $n \to \infty$, $\sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$ converges in distribution to $N(0, I_d)$ with probability 1.*

Proof. We show a highlight of the proof of Lemma 1, in five steps.

Step 1 (Hájek, 1960): Show that simple random sampling and Poisson sampling approaches produce the same limiting distributions.

In the notation of Hájek (1960), suppose we have an infinite sequence of simple random sample experiments indexed by ν. Experiment ν draws a simple random sample of size n_ν from a population of size N_ν given by $\{Y_{\nu 1}, \ldots, Y_{\nu N_\nu}\}$. We assume that $n_\nu \to \infty$ and $N_\nu - n_\nu \to \infty$. In the simple random sampling set-up, a subset s_k of indices $\{1, \ldots, N_\nu\}$ is chosen with probability

$$
P(s_k) = \begin{cases} \binom{N_\nu}{n_\nu}^{-1} & : \quad |s_k| = n_\nu \\ 0 & : \quad \text{else.} \end{cases}
$$

In contrast, in a Poisson sampling approach with mean sample size n_ν, a subset s_k is chosen with probability

$$P(s_k) = \left(\frac{n_\nu}{N_\nu}\right)^k \left(1 - \frac{n_\nu}{N_\nu}\right)^{N_\nu - k}.$$

We say that each experiment produces a simple random sample (SRS) s_n and a Poisson sample s_k such that $s_n \subseteq s_k$ or $s_k \subseteq s_n$. To construct these samples, we take two steps:

(i) Draw $k \sim \text{Binom}(N_\nu, n_\nu/N_\nu)$.
(ii) If $k = n$, choose SRS s_n, and set $s_k = s_n$.
If $k > n$, choose SRS s_k, and then let s_n be an SRS of size n from s_k.
If $k < n$, choose SRS s_n, and then let s_k be an SRS of size k from s_n.

Using the two samples, we define two random variables:

$$\eta_\nu = \sum_{i \in s_n} (Y_{\nu i} - \overline{Y}_\nu) \qquad \text{and} \qquad \eta_\nu^* = \sum_{i \in s_k} (Y_{\nu i} - \overline{Y}_\nu).$$

We can show that the variance of η_ν^* is

$$D\eta_\nu^* = \text{var}(\eta_\nu^*) = \frac{n_\nu}{N_\nu}\left(1 - \frac{n_\nu}{N_\nu}\right) \sum_{i=1}^{N_\nu} (Y_{\nu i} - \overline{Y}_\nu)^2.$$

Under the assumption that $n_\nu \to \infty$ and $N - n_\nu \to \infty$, we can then show that

$$\lim_{\nu \to \infty} \frac{\mathbb{E}[(\eta_\nu - \eta_\nu^*)^2]}{D\eta_\nu^*} = 0. \tag{S1}$$

Remark 2.1 of Hájek (1960) states that (S1) implies that the limiting distributions of $\eta_\nu/\sqrt{D\eta_\nu^*}$ and $\eta_\nu^*/\sqrt{D\eta_\nu^*}$ are the same if they exist, and they exist under the same conditions. To see this, we use Chebyshev's inequality. For $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{\eta_\nu}{\sqrt{D\eta_\nu^*}} - \frac{\eta_\nu^*}{\sqrt{D\eta_\nu^*}}\right| > \epsilon\right) \leq \frac{1}{\epsilon^2}\text{var}\left(\frac{\eta_\nu - \eta_\nu^*}{\sqrt{D\eta_\nu^*}}\right) = \frac{1}{\epsilon^2}\frac{\mathbb{E}[(\eta_\nu - \eta_\nu^*)^2]}{D\eta_\nu^*} \overset{\nu\to\infty}{\to} 0.$$

This means that $\left|\eta_\nu/\sqrt{D\eta_\nu^*} - \eta_\nu^*/\sqrt{D\eta_\nu^*}\right| \overset{p}{\to} 0$. Under this condition, for any distribution W, $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow W$ if and only if $\eta_\nu^*/\sqrt{D\eta_\nu^*} \rightsquigarrow W$.

Since η_ν^* is a sum of independent random variables, it will be easier to work with $\eta_\nu^*/D\eta_\nu^*$ than to work with $\eta_\nu/D\eta_\nu^*$.

Step 2 (Hájek, 1960): Find conditions such that $\eta_\nu/\sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$. (We can think of η_ν as $(n/2)(\overline{Y}_{0,n} - \overline{Y}_n)$ and $D\eta_\nu^*$ as $\text{var}(\sum_{i=1}^n B_i(Y_{ni} - \overline{Y}_n))$ for $B_i \overset{iid}{\sim} \text{Bernoulli}(1/2)$.)

Theorem 3.1 in Hájek (1960) is the key result for asymptotic normality. We present an intermediate result from the proof of Theorem 3.1.

Let $\xi_\nu = \sum_{i \in s_{n,\nu}} Y_{\nu,i}$. (So $\eta_\nu = \xi_\nu - n_\nu \overline{Y}_\nu$.) Let $D\xi_\nu$ be the variance of ξ_ν. Let $S_{\nu\tau}$ be the subset of $S_\nu = \{1, \ldots, N_\nu\}$ on which the inequality

$$|Y_{\nu i} - \overline{Y}_\nu| > \tau\sqrt{D\xi_\nu}$$

holds. Suppose that $n_\nu \to \infty$ and $N_\nu - n_\nu \to \infty$. If

$$\lim_{\nu\to\infty} \frac{\sum_{i\in S_{\nu\tau}} (Y_{\nu i} - \bar{Y}_\nu)^2}{\sum_{i\in S_\nu} (Y_{\nu i} - \bar{Y}_\nu)^2} = 0 \quad \textit{for any } \tau > 0, \tag{S2}$$

then $\eta_\nu / \sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$.

We will show that $\eta_\nu^* / \sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$, and then we can appeal to Step 1's result. η_ν^* is the centered sum of the Poisson sampling terms. We can write η_ν^* as

$$\eta_\nu^* = \sum_{i=1}^{N_\nu} \zeta_{\nu i}, \text{ where } \zeta_{\nu i} = \begin{cases} Y_{\nu i} - \bar{Y}_\nu & \text{with probabilty } n_\nu/N_\nu \\ 0 & \text{with probabilty } 1 - n_\nu/N_\nu. \end{cases}$$

In this setting, Lindeberg's condition for $\eta_\nu^* / \sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$ is for all $\tau > 0$,

$$\lim_{\nu\to\infty} \frac{1}{D\eta_\nu^*} \sum_{i=1}^{N_\nu} \mathbb{E}\left[(\zeta_{\nu i} - \mathbb{E}[\zeta_{\nu i}])^2 \cdot \mathbb{1}\left(|\zeta_{\nu i} - \mathbb{E}[\zeta_{\nu i}| > \tau\sqrt{D\eta_\nu^*} \right) \right] = 0.$$

We can show that (S2) implies that the Lindeberg condition is satisfied. Since Step 1 implies that the limiting distribution of $\eta_\nu / \sqrt{D\eta_\nu^*}$ must be the same as the limiting distribution of $\eta_\nu^* / \sqrt{D\eta_\nu^*}$, we conclude that $\eta_\nu / \sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$.

Step 3: If $d = 1$, show that $\eta_\nu / \sqrt{D\eta_\nu^*} \rightsquigarrow N(0,1)$ implies $\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n) \rightsquigarrow N(0,1)$.

This is mostly a matter of adapting Step 2's result to our setting. When $n_\nu / N_\nu = 1/2$, η_ν is the same random variable as $(n/2)(\bar{Y}_{0,n} - \bar{Y}_n)$. Using the formula for $D\eta_\nu^*$,

$$\frac{\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n)}{\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_{ni} - \bar{Y}_n)^2}} = \frac{(n/2)(\bar{Y}_{0,n} - \bar{Y}_n)}{\sqrt{\frac{1}{4}\sum_{i=1}^n (Y_{ni} - \bar{Y}_n)^2}} \overset{d}{=} \frac{\eta_\nu}{\sqrt{D\eta_\nu^*}} \rightsquigarrow N(0,1).$$

In addition, $\sqrt{\frac{1}{n}\sum_{i=1}^n (Y_{ni} - \bar{Y}_n)^2} / \sqrt{\text{var}(Y_{ni})} \xrightarrow{p} 1$. By Slutsky's Theorem, $\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n) \rightsquigarrow N(0,1)$.

Step 4 (Li & Ding, 2017): If $Y_{n1}, \dots, Y_{nn} \sim N(\theta^*, 1)$, show that the condition of Step 2 is satisfied with probability 1.

These results come from page 2 of the appendix of Li & Ding (2017). The authors show that if the Y_{ni}s are iid draws from a superpopulation with $2 + \epsilon$ ($\epsilon > 0$) absolute moments and nonzero variance, then $(1/n)\max_{1\le i\le n}(Y_{ni} - \bar{Y}_n)^2 \equiv m_n/n \to 0$ with probability 1. Furthermore, they show that $m_n/n \to 0$ implies their condition (A2), which is a rewriting of Hájek (1960)'s condition (S2).

Since $N(\theta^*, 1)$ satisfies the superpopulation conditions, condition (S2) is satisfied with probability 1. Then following Steps 2 and 3, $\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n) \rightsquigarrow N(0,1)$.

Step 5 (Hájek, 1960): Extend results to $d > 1$.

In d dimensions, suppose $Y_{n1}, \dots, Y_{nn} \sim N(\theta^*, I_d)$. Remark 3.2 of Hájek (1960) notes that we can user the Cramér-Wold device to extend the results to the multivariate case. Let $Z = (Z^{(1)}, \dots, Z^{(d)})$ represent the $N(0, I_d)$ distribution. Then for each component, $Z^{(j)} \sim N(0,1)$. By the Cramér-Wold device, we can say that $\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n) \rightsquigarrow Z$ if and only if for any $\lambda \in \mathbb{R}^d$, $\sum_{j=1}^d \lambda^{(j)} \sqrt{n}(\bar{Y}_{0,n}^{(j)} - \bar{Y}_n^{(j)}) \rightsquigarrow \sum_{j=1}^d \lambda^{(j)} Z^{(j)}$.

For any dimension j, we can think of $Y_{n1}^{(j)}, \dots, Y_{nn}^{(j)}$ as draws from a $N(\theta^{*(j)}, 1)$ superpopulation. So the superpopulation conditions from Step 4 are satisfied, which means $\sqrt{n}(\bar{Y}_{0,n}^{(j)} - \bar{Y}_n^{(j)}) \rightsquigarrow Z^{(j)}$. We conclude that $\sqrt{n}(\bar{Y}_{0,n} - \bar{Y}_n) \rightsquigarrow N(0, I_d)$. □

Lemma 2. *Assume* $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$ *is a sequence of data sets such that* $\mathcal{D}_n = \{Y_{n1}, Y_{n2}, \ldots, Y_{nn}\}$ *with observations* $Y_{nj} \overset{iid}{\sim} N(\theta^*, I_d)$*. Let* $\mathcal{D}_{0,n}$ *be a sample of* $n/2$ *observations from* $\mathcal{D}_n$*. Define* $\overline{Y}_n = (1/n)\sum_{i=1}^n Y_{ni}$ *and* $\overline{Y}_{0,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{0,n}} Y_{ni}$*. Let* $c > 0$*, and let* (θ_n) *be a sequence that satisfies* $\|\overline{Y}_n - \theta_n\| \le c/\sqrt{n}$ *for all* n*. Define* $X_n \equiv \sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$*. Let* Z *denote a* $N(0, I_d)$ *random variable. Then*

$$\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] - \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^T Z + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] = o_P(1).$$

Proof. Since (θ_n) is chosen such that $\|\overline{Y}_n - \theta_n\| \le c/\sqrt{n}$, we can re-write $\theta_n = \overline{Y}_n + (c/\sqrt{n})v_n$, where $v_n \in \mathbb{R}^d$ satisfies $\|v_n\| \le 1$ for all n.

Define a function f by

$$f(x_n, v_n) \equiv \exp\left(-\frac{3}{4}x_n^T x_n - \frac{c}{2}x_n^T v_n\right).$$

f is clearly a continuous function. We can also show that f is bounded. Define

$$g(x_n, v_n) \equiv -\frac{3}{4}x_n^T x_n - \frac{c}{2}x_n^T v_n$$

so that $f(x_n, v_n) = \exp(g(x_n, v_n))$. We can see that

$$\frac{\partial}{\partial x_n}g(x_n, v_n) = -\frac{3}{2}x_n - \frac{c}{2}v_n \overset{set}{=} \vec{0}$$

is solved by $x_n = -(c/3)v_n$. Since $g(x_n, v_n)$ is concave in x_n, $g(x_n, v_n)$ is maximized at $x_n = -(c/3)v_n$ for any v_n. Since $f(x_n, v_n) = \exp(g(x_n, v_n))$, $f(x_n, v_n)$ is also maximized at this value of x_n for any v_n. Under the assumption that $\|v_n\| \le 1$, we see

$$\begin{aligned} f(x_n, v_n) &\le \exp\left(-\frac{3}{4}\left(-\frac{c}{3}\right)^2 v_n^T v_n - \frac{c}{2}\left(-\frac{c}{3}\right)v_n^T v_n\right) \\ &= \exp\left(-\frac{c^2}{12}\|v_n\|^2 + \frac{c^2}{6}\|v_n\|^2\right) \\ &\le \exp\left(\frac{c^2}{12}\right). \end{aligned}$$

Thus, $f(x_n, v_n)$ is a continuous and bounded function.

The claim of Lemma 2 is equivalent to $\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] = o_P(1)$. The Portmanteau Theorem provides several equivalent definitions of convergence in distribution, including that $X_n \rightsquigarrow Z$ if and only if $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(Z)]$ for every continuous, bounded function h. We prove the result on $f(X_n, v_n)$ by modifying the Van der Vaart (2000), Chapter 2, proof of this Portmanteau Theorem result.

Let $\gamma > 0$. Fix $\epsilon > 0$ such that

$$\epsilon < \gamma \,/\, (3 + 3\exp(c^2/12)). \tag{S3}$$

Choose a large enough compact rectangle I such that

$$\mathbb{P}(Z \notin I) < \epsilon. \tag{S4}$$

Let $\mathcal{B}_1(0)$ be the d-dimensional ball of radius 1 centered at 0. By construction, each $v_n \in \mathcal{B}_1(0)$. Since f is continuous and $I \times \mathcal{B}_1(0)$ is compact, $f(x_n, v_n)$ is uniformly continuous on $I \times \mathcal{B}_1(0)$. We can thus partition $I \times \mathcal{B}_1(0)$ into J compact regions $I_j \times V_j$ where $I \times \mathcal{B}_1(0) = \cup_{j=1}^J (I_j \times V_j)$ such that for any j and for any $(x_{n1}, v_{n1}), (x_{n2}, v_{n2}) \in I_j \times V_j$, $|f(x_{n1}, v_{n1}) - f(x_{n2}, v_{n2})| < \epsilon$. (For instance the I_j regions may be rectangles and the V_j regions may be rectangles truncated at the boundaries of $\mathcal{B}_1(0)$. These rectangular regions may be appropriately sized such that within a region $I_j \times V_j$, $d((x_{n1}, v_{n1}), (x_{n2}, v_{n2}))$ is small enough that $|f(x_{n1}, v_{n1}) - f(x_{n2}, v_{n2})| < \epsilon$.)

Select a point (x'_j, v'_j) from each $I_j \times V_j$. Define

$$f_\epsilon(x, v) = \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{1}((x, v) \in I_j \times V_j).$$

For a given sample $\mathcal{D}_n$, we note that there are $\binom{n}{n/2}$ possible values of X_n, since there are $\binom{n}{n/2}$ possible values of $\overline{Y}_{0,n}$. We denote the sum over all possible values of X_n as $\sum_{X_n}$.

Note that

$$\begin{aligned}
&|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]| \\
&= \left| \binom{n}{n/2}^{-1} \sum_{X_n} f(X_n, v_n) - \binom{n}{n/2}^{-1} \sum_{X_n} f_\epsilon(X_n, v_n) \right| \\
&= \left| \binom{n}{n/2}^{-1} \sum_{X_n} \big[(f(X_n, v_n) - f_\epsilon(X_n, v_n))\mathbb{1}(X_n \in I) + (f(X_n, v_n) - f_\epsilon(X_n, v_n))\mathbb{1}(X_n \notin I) \big] \right| \\
&\leq \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)|\mathbb{1}(X_n \in I) + \\
&\quad \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)|\mathbb{1}(X_n \notin I) \\
&= \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)|\mathbb{1}(X_n \in I, v_n \in \mathcal{B}_1(0)) + \\
&\quad \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n) - f_\epsilon(X_n, v_n)|\mathbb{1}(X_n \notin I) \\
&< \binom{n}{n/2}^{-1} \sum_{X_n} \epsilon + \binom{n}{n/2}^{-1} \sum_{X_n} |f(X_n, v_n)|\mathbb{1}(X_n \notin I) \\
&\leq \epsilon + \exp\left(c^2/12\right) \mathbb{P}(X_n \notin I \mid \mathcal{D}_n). \qquad \text{(S5)}
\end{aligned}$$

Similarly, we show that

$$
\begin{aligned}
&\Big|\mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\Big| \\
&= \Big|\mathbb{E}\big[(f(Z, v_n) - f_\epsilon(Z, v_n))\mathbb{1}(Z \in I) + (f(Z, v_n) - f_\epsilon(Z, v_n))\mathbb{1}(Z \notin I)\big]\Big| \\
&\leq \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big|\mathbb{1}(Z \in I) \mid \mathcal{D}_n\right] + \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big|\mathbb{1}(Z \notin I) \mid \mathcal{D}_n\right] \\
&= \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big|\mathbb{1}(Z \in I, v_n \in \mathcal{B}_1(0)) \mid \mathcal{D}_n\right] + \mathbb{E}\left[\Big|f(Z, v_n) - f_\epsilon(Z, v_n)\Big|\mathbb{1}(Z \notin I) \mid \mathcal{D}_n\right] \\
&< \epsilon + \exp(c^2/12)\mathbb{P}(Z \notin I \mid \mathcal{D}_n) \\
&= \epsilon + \exp(c^2/12)\mathbb{P}(Z \notin I) \\
&< \epsilon + \epsilon \exp(c^2/12). \qquad \text{(S6)}
\end{aligned}
$$

In addition, we see that

$$
\begin{aligned}
&\left|\mathbb{E}\left[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n\right] - \mathbb{E}\left[f_\epsilon(Z, v_n) \mid \mathcal{D}_n\right]\right| \\
&= \left|\binom{n}{n/2}^{-1} \sum_{X_n} f_\epsilon(X_n, v_n) - \mathbb{E}[f_\epsilon(Z, v_n)]\right| \\
&= \left|\binom{n}{n/2}^{-1} \sum_{X_n} \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{1}((X_n, v_n) \in I_j \times V_j) - \sum_{j=1}^{J} f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\mathbb{1}(v_n \in V_j)\right| \\
&\leq \sum_{j=1}^{J} \left|\binom{n}{n/2}^{-1} \sum_{X_n} f(x'_j, v'_j)\mathbb{1}(X_n \in I_j)\mathbb{1}(v_n \in V_j) - f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\mathbb{1}(v_n \in V_j)\right| \\
&\leq \sum_{j=1}^{J} \left|\binom{n}{n/2}^{-1} \sum_{X_n} f(x'_j, v'_j)\mathbb{1}(X_n \in I_j) - f(x'_j, v'_j)\mathbb{P}(Z \in I_j)\right| \\
&= \sum_{j=1}^{J} \left|f(x'_j, v'_j) \left[\binom{n}{n/2}^{-1} \sum_{X_n} \mathbb{1}(X_n \in I_j) - \mathbb{P}(Z \in I_j)\right]\right| \\
&\leq \sum_{j=1}^{J} \left|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\right| \times \left|f(x'_j, v'_j)\right|. \qquad \text{(S7)}
\end{aligned}
$$

For the sequence of datasets $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$, Lemma 1 establishes that $X_n \rightsquigarrow N(0, I_d)$ with probability 1. This tells us that with probability 1 over the randomness in sequences $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$, $\lim_{n\to\infty} \mathbb{P}(X_n \in I \mid \mathcal{D}_n) = \mathbb{P}(Z \in I)$. Since almost sure convergence implies convergence in probability, for any $\delta > 0$,

$$
\lim_{n\to\infty} \mathbb{P}\big(|\mathbb{P}(X_n \in I \mid \mathcal{D}_n) - \mathbb{P}(Z \in I)| > \delta\big) = 0 \qquad \text{(S8)}
$$

$$
\text{and } \lim_{n\to\infty} \mathbb{P}\big(|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)| > \delta\big) = 0 \text{ for } 1 \leq j \leq J. \qquad \text{(S9)}
$$

The outer probability is over the randomness in the sequences $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$.

Now we see

$$
\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma\big) \\
&\le \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]\big| + \\
&\qquad\qquad \big|\mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\big| + \\
&\qquad\qquad \big|\mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma\big) \\
&\le \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f_\epsilon(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f_\epsilon(Z, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma/3\big) \\
&\le \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) > \gamma/3\big) + \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \epsilon\exp(c^2/12) > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\left(\sum_{j=1}^{J} \big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| \times |f(x_j', v_j')| > \gamma/3\right) \text{ by (S5), (S6), and (S7)} \\
&= \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) > \gamma/3\big) + \\
&\quad \lim_{n\to\infty} \mathbb{P}\left(\sum_{j=1}^{J} \big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| \times |f(x_j', v_j')| > \gamma/3\right) \text{ by (S3)} \\
&\le \lim_{n\to\infty} \mathbb{P}\big(\epsilon + \exp(c^2/12)\left(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I)\right) > \gamma/3 - \exp(c^2/12)\mathbb{P}(Z \notin I)\big) + \\
&\quad \lim_{n\to\infty} \sum_{j=1}^{J} \mathbb{P}\left(\big|\mathbb{P}(X_n \in I_j \mid \mathcal{D}_n) - \mathbb{P}(Z \in I_j)\big| > (\gamma/3)|f(x_j', v_j)|^{-1}\right) \\
&\le \lim_{n\to\infty} \mathbb{P}\left(\epsilon + \exp(c^2/12)(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I)) > \gamma/3 - \epsilon\exp(c^2/12)\right) \text{ by (S4) and (S9)} \\
&= \lim_{n\to\infty} \mathbb{P}\left(\mathbb{P}(X_n \notin I \mid \mathcal{D}_n) - \mathbb{P}(Z \notin I) > \frac{\gamma - 3\epsilon - 3\epsilon\exp(c^2/12)}{3\exp(c^2/12)}\right) \\
&= 0 \text{ by (S3) and (S8).}
\end{aligned}
$$

We have shown that for arbitrary $\gamma > 0$,

$$\lim_{n\to\infty} \mathbb{P}\big(\big|\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n]\big| > \gamma\big) = 0.$$

We conclude that $\mathbb{E}[f(X_n, v_n) \mid \mathcal{D}_n] - \mathbb{E}[f(Z, v_n) \mid \mathcal{D}_n] = o_P(1)$. □

Theorem 2. *Assume we have a sequence of datasets $(\mathcal{D}_n)_{n\in 2\mathbb{N}}$, where $\mathcal{D}_n = \{Y_{n1}, \ldots, Y_{nn}\}$ and each Y_{ni} is an independent observation from $N(\theta^*, I_d)$. Let $\mathcal{D}_{0,n}$ be a sample of $n/2$ observations from $\mathcal{D}_n$, and let $\mathcal{D}_{1,n} = \mathcal{D}_n \backslash \mathcal{D}_{0,n}$. Define $\overline{Y}_n = (1/n)\sum_{i=1}^n Y_{ni}$, $\overline{Y}_{0,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{0,n}} Y_{ni}$, and $\overline{Y}_{1,n} = (2/n)\sum_{Y_{ni}\in\mathcal{D}_{1,n}} Y_{ni}$. Let $c > 0$, and let (θ_n) be a sequence that satisfies $\|\overline{Y}_n - \theta_n\| \le c/\sqrt{n}$ for all n. Then*

$$\mathbb{E}\{T_n(\theta_n) \mid \mathcal{D}_n\} \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} = 1 + o_P(1).$$

Proof. Define $X_n \equiv \sqrt{n}(\overline{Y}_{0,n} - \overline{Y}_n)$ and let $Z \sim N(0, I_d)$. In addition, define $\mu_n \equiv (\sqrt{n}/5)(\overline{Y}_n - \theta_n)$ and $\Omega \equiv (2/5)I_d$. Then

$$
\begin{aligned}
&\mathbb{E}[T_n(\theta_n) \mid \mathcal{D}_n] \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} \\
&= \mathbb{E}\left[\exp\left(-\frac{n}{4}\|\overline{Y}_{0,n} - \overline{Y}_{1,n}\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \;/\; \left\{\exp\left(\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}\right\} \\
&= \mathbb{E}\left[\exp\left(-\frac{n}{4}\|2\overline{Y}_{0,n} - 2\overline{Y}_n\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-n\|\overline{Y}_{0,n} - \overline{Y}_n\|^2 + \frac{n}{4}\|\overline{Y}_{0,n} - \overline{Y}_n + \overline{Y}_n - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3n}{4}\|\overline{Y}_{0,n} - \overline{Y}_n\|^2 + \frac{n}{2}(\overline{Y}_{0,n} - \overline{Y}_n)^T(\overline{Y}_n - \theta_n) + \frac{n}{4}\|\overline{Y}_n - \theta_n\|^2\right) \mid \mathcal{D}_n\right] \times \\
&\qquad \exp\left(-\frac{3n}{10}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] \exp\left(-\frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{-d/2} \\
&= \mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^T X_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] / \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^T Z + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] \qquad \text{(S10)} \\
&= 1 + o_P(1). \qquad \text{(S11)}
\end{aligned}
$$

Step (S10) holds because

$$
\begin{aligned}
&\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^T Z + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] \\
&= \int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}|I_d|^{1/2}}\exp\left(-\frac{1}{2}z^T z\right)\exp\left(-\frac{3}{4}z^T z + \frac{\sqrt{n}}{2}z^T\left(\overline{Y}_n - \theta_n\right)\right)\right]dz \\
&= \int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}}\exp\left(-\frac{5}{4}z^T z + \frac{\sqrt{n}}{2}z^T\left(\overline{Y}_n - \theta_n\right)\right)\right]dz \\
&= |\Omega|^{1/2}\int_{\mathbb{R}^d}\left[\frac{1}{(2\pi)^{d/2}|\Omega|^{1/2}}\exp\left(-\frac{1}{2}(z-\mu_n)^T\Omega^{-1}(z-\mu_n) + \frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)\right]dz \qquad \text{(S12)} \\
&= \exp\left(\frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)|\Omega|^{1/2} \\
&= \exp\left(\frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}.
\end{aligned}
$$

Step (S12) uses the following equality:

$$\begin{aligned}
&-\frac{5}{4}z^Tz + \frac{\sqrt{n}}{2}z^T(\overline{Y}_n - \theta_n) \\
&\quad = -\frac{5}{4}\left[z^Tz - \frac{2\sqrt{n}}{5}z^T(\overline{Y}_n - \theta_n) + \frac{n}{25}(\overline{Y}_n - \theta_n)^T(\overline{Y}_n - \theta_n) - \frac{n}{25}(\overline{Y}_n - \theta_n)^T(\overline{Y}_n - \theta_n)\right] \\
&\quad = -\frac{5}{4}\left(z - \frac{\sqrt{n}}{5}(\overline{Y}_n - \theta_n)\right)^T\left(z - \frac{\sqrt{n}}{5}(\overline{Y}_n - \theta_n)\right) + \frac{n}{20}\|\overline{Y}_n - \theta_n\|^2 \\
&\quad = -\frac{1}{2}\left(z - \frac{\sqrt{n}}{5}(\overline{Y}_n - \theta_n)\right)^T\left(\frac{5}{2}I_d\right)\left(z - \frac{\sqrt{n}}{5}(\overline{Y}_n - \theta_n)\right) + \frac{n}{20}\|\overline{Y}_n - \theta_n\|^2 \\
&\quad = -\frac{1}{2}(z - \mu_n)^T\Omega^{-1}(z - \mu_n) + \frac{n}{20}\|\overline{Y}_n - \theta_n\|^2.
\end{aligned}$$

To justify step (S11), note that $\mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right]$, which equals $\exp\left(\frac{n}{20}\|\overline{Y}_n - \theta_n\|^2\right)\left(\frac{2}{5}\right)^{d/2}$, is bounded between $(2/5)^{d/2}$ and $\exp(c^2/20)(2/5)^{d/2}$ under the assumption that $\|\overline{Y}_n - \theta_n\| \leq c/\sqrt{n}$. By Lemma 2,

$$\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^TX_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] - \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] = o_P(1).$$

Combining these two facts, we conclude that

$$\mathbb{E}\left[\exp\left(-\frac{3}{4}X_n^TX_n + \frac{\sqrt{n}}{2}X_n^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] / \mathbb{E}\left[\exp\left(-\frac{3}{4}Z^TZ + \frac{\sqrt{n}}{2}Z^T\left(\overline{Y}_n - \theta_n\right)\right) \mid \mathcal{D}_n\right] = 1 + o_P(1).$$

□

Theorem 3. *Let $Y_1, \ldots, Y_n \sim N(\theta^*, I_d)$. The splitting proportion that minimizes $\mathbb{E}[r^2\{C_n^{split}(\alpha)\}]$ is*

$$p_0^* = 1 - \frac{\sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)} - 2d}{4\log\left(\frac{1}{\alpha}\right)}.$$

Proof. Recall that p_0 represents the proportion of observations that we place in $\mathcal{D}_0$.

We know that

$$\overline{Y}_0 \sim N\left(\theta^*,\ Var = \frac{1}{np_0}I_d\right)$$
$$\overline{Y}_1 \sim N\left(\theta^*,\ Var = \frac{1}{n(1-p_0)}I_d\right)$$

Since all observations in $\mathcal{D}_0$ and $\mathcal{D}_1$ are mutually independent, this implies

$$\overline{Y}_0 - \overline{Y}_1 \sim N\left(0, \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)I_d\right) \tag{S13}$$

and, hence,

$$\left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)^{-1/2}\left(\overline{Y}_0 - \overline{Y}_1\right) \sim N\left(0, I_d\right).$$

We now see

$$\|\overline{Y}_0 - \overline{Y}_1\|^2 = \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) \left\| \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right)^{-1/2} (\overline{Y}_0 - \overline{Y}_1) \right\|^2$$
$$\stackrel{d}{=} \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) \chi^2_d. \tag{S14}$$

When $p_0 = \frac{1}{2}$, this expression is $\frac{4}{n}\chi^2_d$, as shown in the derivation of equation 7.

Setting $\widehat{\theta}_1 = \overline{Y}_1$, at $\theta \in \mathbb{R}^d$ we construct the test statistic:

$$\begin{aligned} T_n(\theta) &= \frac{\prod_{Y_{0i}\in D_0} \exp\left(-\frac{1}{2}(Y_{0i} - \widehat{\theta}_1)^T (Y_{0i} - \widehat{\theta}_1)\right)}{\prod_{Y_{0i}\in D_0} \exp\left(-\frac{1}{2}(Y_{0i} - \theta)^T (Y_{0i} - \theta)\right)} \\ &= \exp\left(\sum_{Y_{0i}\in D_0} \left(-\frac{1}{2}(\overline{Y}_0 - \overline{Y}_1)^T (\overline{Y}_0 - \overline{Y}_1) + \frac{1}{2}(\overline{Y}_0 - \theta)^T (\overline{Y}_0 - \theta) \right) \right) \\ &= \exp\left(-\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2\right) \end{aligned}$$

Using a split proportion of p_0, the split LRT confidence set is now

$$\begin{aligned} C_n^{\text{split}} &= \left\{ \theta \in \Theta : \exp\left(-\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2\right) \le \frac{1}{\alpha} \right\} \\ &= \left\{ \theta \in \Theta : -\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2 \le \log\left(\frac{1}{\alpha}\right) \right\} \\ &= \left\{ \theta \in \Theta : \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2 \le \log\left(\frac{1}{\alpha}\right) + \frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 \right\} \\ &= \left\{ \theta \in \Theta : \|\overline{Y}_0 - \theta\|^2 \le \frac{2}{np_0}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2 \right\} \end{aligned}$$

The squared radius is thus $R^2(C_n^{\text{split}}) = \frac{2}{np_0}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2$. By (S14), the expected squared radius at a given value of p_0 is

$$r(p_0) = \frac{2}{np_0}\log\left(\frac{1}{\alpha}\right) + \left(\frac{1}{np_0} + \frac{1}{n(1-p_0)}\right) d.$$

We can now minimize this function:

$$
\begin{aligned}
0 &\overset{set}{=} \frac{\partial}{\partial p_0} r(p_0) = \frac{-2}{np_0^2}\log\left(\frac{1}{\alpha}\right) - \frac{d}{np_0^2} + \frac{d}{n(1-p_0)^2} \\
&\Updownarrow \\
0 &= -2(1-p_0)^2 \log\left(\frac{1}{\alpha}\right) - d(1-p_0)^2 + dp_0^2 \\
&= -2(1-2p_0+p_0^2)\log\left(\frac{1}{\alpha}\right) - d(1-2p_0+p_0^2) + dp_0^2 \\
&= -2\log\left(\frac{1}{\alpha}\right) + 4p_0\log\left(\frac{1}{\alpha}\right) - 2p_0^2\log\left(\frac{1}{\alpha}\right) - d + 2dp_0 - dp_0^2 + dp_0^2 \\
&= p_0^2\left(-2\log\left(\frac{1}{\alpha}\right)\right) + p_0\left(4\log\left(\frac{1}{\alpha}\right) + 2d\right) + \left(-2\log\left(\frac{1}{\alpha}\right) - d\right).
\end{aligned}
$$

This is now a quadratic expression in p_0. Thus, this formula is solved by

$$
\begin{aligned}
p_0 &= \frac{-4\log\left(\frac{1}{\alpha}\right) - 2d \pm \sqrt{\left(4\log\left(\frac{1}{\alpha}\right) + 2d\right)^2 - 4\left(-2\log\left(\frac{1}{\alpha}\right)\right)\left(-2\log\left(\frac{1}{\alpha}\right) - d\right)}}{2\left(-2\log\left(\frac{1}{\alpha}\right)\right)} \\
&= \frac{4\log\left(\frac{1}{\alpha}\right) + 2d \pm \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)}.
\end{aligned}
$$

We now consider the $\pm$ choice. In the $+$ direction, we have

$$
p_0 = \frac{4\log\left(\frac{1}{\alpha}\right) + 2d + \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)} = 1 + \frac{2d + \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)} > 1.
$$

However, in the $-$ direction, we can show that $p_0 \in \left(\frac{1}{2}, 1\right)$. We note that

$$
\begin{aligned}
2d &< \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)} < \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right) + 4\left(\log\left(\frac{1}{\alpha}\right)\right)^2} \\
&= \sqrt{\left(2d + 2\log\left(\frac{1}{\alpha}\right)\right)^2} = 2d + 2\log\left(\frac{1}{\alpha}\right).
\end{aligned}
$$

So

$$
p_0 = 1 + \frac{2d - \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)} < 1 + \frac{2d - 2d}{4\log\left(\frac{1}{\alpha}\right)} = 1
$$

and

$$
p_0 = 1 + \frac{2d - \sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)}}{4\log\left(\frac{1}{\alpha}\right)} > 1 + \frac{2d - 2d - 2\log\left(\frac{1}{\alpha}\right)}{4\log\left(\frac{1}{\alpha}\right)} = 1 - \frac{1}{2} = \frac{1}{2}.
$$

This means that

$$
p_0^* = 1 - \frac{\sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)} - 2d}{4\log\left(\frac{1}{\alpha}\right)}
$$

optimizes $r(p_0)$, and $p_0^* \in \left(\frac{1}{2}, 1\right)$. Furthermore, this optimum must be a minimum, since for any $p_0 \in (0,1)$,

$$\frac{\partial^2}{\partial p_0^2} r(p_0) = \frac{4}{np_0^3} \log\left(\frac{1}{\alpha}\right) + \frac{2d}{np_0^3} + \frac{2d}{n(1-p_0)^3} > 0.$$

We can use L'Hôpital's Rule to show that $p_0^* \to \frac{1}{2}$ as $d \to \infty$:

$$\begin{aligned}
\lim_{d\to\infty} p_0^* &= 1 - \lim_{d\to\infty} \frac{\sqrt{4d^2 + 8d\log\left(\frac{1}{\alpha}\right)} - 2d}{4\log\left(\frac{1}{\alpha}\right)} \\
&= 1 - \lim_{d\to\infty} \frac{\sqrt{4 + (8/d)\log(1/\alpha)} - 2}{(4/d)\log(1/\alpha)} \\
&= 1 - \lim_{d\to\infty} \frac{\frac{1}{2}\left(4 + (8/d)\log(1/\alpha)\right)^{-1/2}(-8/d^2)\log(1/\alpha)}{(-4/d^2)\log(1/\alpha)} \\
&= 1 - \lim_{d\to\infty} \left(4 + (8/d)\log(1/\alpha)\right)^{-1/2} \\
&= \frac{1}{2}.
\end{aligned}$$

We conclude that as $d \to \infty$ for fixed α, the optimal choice of $p_0^* \to 0.5$. □

Theorem 4. *Suppose $Y_1, \ldots, Y_n$ are iid observations from $N(\theta^*, I_d)$. Split the sample such that $\mathcal{D}_0$ and $\mathcal{D}_1$ each contain $\frac{n}{2}$ observations. Use $\mathcal{D}_0$ and $\mathcal{D}_1$ to define the split and cross-fit sets. Then Volume$\{C_n^{CF}(\alpha)\} \leq$ Volume$\{C_n^{split}(\alpha)\}$. Equality holds only when $\overline{Y}_0 = \overline{Y}_1$.*

Proof. Let $\theta \in C_n^{\text{CF}}(\alpha)$. Then

$$\begin{aligned}
&\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y} - \theta\|^2\right) \\
&= \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\left\|\frac{1}{2}(\overline{Y}_0 - \theta) + \frac{1}{2}(\overline{Y}_1 - \theta)\right\|^2\right) \\
&\leq \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_0 - \theta\|^2 + \frac{n}{8}\|\overline{Y}_1 - \theta\|^2\right) \qquad \text{(S15)} \\
&= \exp\left(-\frac{n}{8}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_0 - \theta\|^2 - \frac{n}{8}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{8}\|\overline{Y}_1 - \theta\|^2\right) \\
&\leq \frac{1}{2}\Bigg[\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \theta\|^2\right) + \\
&\qquad \exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_1 - \theta\|^2\right)\Bigg] \qquad \text{(S16)} \\
&< \frac{1}{\alpha}.
\end{aligned}$$

Line (S15) holds because $\|\cdot\|^2$ is convex. Line (S16) holds because $\exp(\cdot)$ is convex. Thus, $C_n^{\text{CF}}(\alpha) \subseteq \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < \frac{4}{n}\log(\frac{1}{\alpha}) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$, which has the same volume as $C_n^{\text{split}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y}_0 - \theta\|^2 < \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$. Hence, $Vol\left(C_n^{\text{CF}}(\alpha)\right) \leq Vol\left(C_n^{\text{split}}(\alpha)\right)$.

Furthermore, since $\|\cdot\|^2$ and $\exp(\cdot)$ are strictly convex, equality holds in (S15) and (S16) only when $\overline{Y}_0 = \overline{Y}_1$. If $\overline{Y}_0 = \overline{Y}_1$, then $C_n^{\text{CF}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < rem\frac{4}{n}\log(\frac{1}{\alpha}) + \|\overline{Y}_0 - \overline{Y}_1\|^2\right\}$, which means $Vol\left(C_n^{\text{CF}}(\alpha)\right) = Vol\left(C_n^{\text{split}}(\alpha)\right)$. □

Theorem 5. *Assume $c_{\alpha,d} + \log(\alpha) > d - 2$. Let $f_d(x)$ be the probability density function of the χ_d^2 distribution, and let $c_{\alpha,d}$ be the upper α quantile of the χ_d^2 distribution. Then*

$$\begin{aligned} &\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \geq 1 - \alpha - \log(1/\alpha) f_d\{c_{\alpha,d} + \log(\alpha)\} \\ \textit{and} \quad &\mathbb{P}\left[r^2\{C_n^{split}(\alpha)\}/r^2\{C_n^{LRT}(\alpha)\} \leq 4\right] \leq 1 - \alpha - \log(1/\alpha) f_d(c_{\alpha,d}). \end{aligned}$$

Proof. We use the fact that $r^2(C_n^{\text{split}}(\alpha)) = \frac{4}{n}\log(1/\alpha) + \|\overline{Y}_0 - \overline{Y}_1\|^2$. As established in the proof of Theorem 3 and the derivation of equation 7, we know that $\|\overline{Y}_0 - \overline{Y}_1\|^2 \overset{d}{=} (4/n)\chi_d^2$. Let $X \sim \chi_d^2$. Note that $\log(\alpha) < 0$. Then

$$\begin{aligned} \mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \,/\, r^2(C_n^{\text{LRT}}(\alpha)) \leq 4\right) &= \mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \leq \frac{4}{n} c_{\alpha,d}\right) \\ &= \mathbb{P}\left(\frac{4}{n}\log(1/\alpha) + \frac{4}{n}X \leq \frac{4}{n}c_{\alpha,d}\right) \\ &= \mathbb{P}\left(\log(1/\alpha) + X \leq c_{\alpha,d}\right) \\ &= \mathbb{P}(X \leq c_{\alpha,d} + \log(\alpha)) \\ &= \mathbb{P}(X \leq c_{\alpha,d}) - \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}) \\ &= 1 - \alpha - \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}). \end{aligned}$$

Now we need to bound $\mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d})$. Under the assumed conditions, we show that the χ_d^2 pdf is decreasing on $[c_{\alpha,d} + \log(\alpha), c_{\alpha,d}]$. Let $f_d(x)$ be the χ_d^2 pdf. The following five statements are equivalent:

$$\begin{aligned} 0 &> \frac{\partial}{\partial x} f_d(x) \\ 0 &> \frac{1}{2^{d/2}\,\Gamma(d/2)}\left[\left(\frac{d}{2} - 1\right) x^{d/2-2} e^{-x/2} + x^{d/2-1}\left(-\frac{1}{2}e^{-x/2}\right)\right] \\ x^{d/2-1}\left(\frac{1}{2}e^{-x/2}\right) &> \left(\frac{d}{2} - 1\right) x^{d/2-2} e^{-x/2} \\ \frac{x}{2} &> \frac{d}{2} - 1 \\ x &> d - 2 \end{aligned}$$

By our initial assumption, $c_{\alpha,d} + \log(\alpha) > d - 2$. Thus, $f_d(x)$ is decreasing on $[c_{\alpha,d} + \log(\alpha), c_{\alpha,d}]$. Since the interval has length $\log(1/\alpha)$,

$$\log(1/\alpha) f_d(c_{\alpha,d}) \;\leq\; \mathbb{P}(c_{\alpha,d} + \log(\alpha) \leq X \leq c_{\alpha,d}) \;\leq\; \log(1/\alpha) f_d(c_{\alpha,d} + \log(\alpha)).$$

The bounds on $\mathbb{P}\left(r^2(C_n^{\text{split}}(\alpha)) \,/\, r^2(C_n^{\text{LRT}}(\alpha)) \leq 4\right)$ follow immediately. □

Before proving Theorem 6, we establish Lemma 3 and Lemma 4.

Lemma 3. *Assume the doughnut null test setting. Let $\mathcal{P}_{\Theta_0}$ be the set of all convex combinations of $N(\theta, I_d)$ densities such that $\|\theta\| \in [0.5, 1]$. When $\|\overline{Y}_1\| > 1$ and $\widehat{\theta}_1 = \overline{Y}_1$, the RIPR of $p_{\widehat{\theta}_1}$ onto $\mathcal{P}_{\Theta_0}$ is $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$.*

Proof. Suppose $\|\overline{Y}_1\| > 1$. Defining $\widehat{\theta}_1 = \overline{Y}_1$ as in Table 1, $\|\widehat{\theta}_1\| > 1$. The RIPR of $\widehat{\theta}_1$ onto the convex set $\mathcal{P}_{\Theta_0}$ minimizes $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \| p_0)$ out of all densities $p_0 \in \mathcal{P}_{\Theta_0}$. Suppose $p_0 \in \mathcal{P}_{\Theta_0}$. Then we can write p_0 as a mixture of $N(\theta_k, I_d)$ densities. We write $p_0 = \sum_{k=1}^K w_k p_{\theta_k}$, where $K \in \mathbb{N}$, $\sum_{k=1}^K w_k = 1$, and for each $k = 1, \ldots, K$, $0 < w_k < 1$ and $\|\theta_k\| \in [0.5, 1]$. Note that $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|} \in \mathcal{P}_0$. To prove that $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}) = \inf_{p_0 \in \mathcal{P}_{\Theta_0}} D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_0)$, we show $D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}) \leq D_{\mathrm{KL}}(p_{\widehat{\theta}_1} \parallel \sum_{k=1}^K w_k p_{\theta_k})$.

$$
\begin{aligned}
&D_{\mathrm{KL}}\left(p_{\widehat{\theta}_1} \,\Big|\Big|\, \sum_{k=1}^K w_k p_{\theta_k}\right) - D_{\mathrm{KL}}\left(p_{\widehat{\theta}_1} \parallel p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}\right) \\
&\quad = \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1}(y)}{\sum_{k=1}^K w_k p_{\theta_k}(y)}\right) dy - \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right) dy \\
&\quad = \int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}{\sum_{k=1}^K w_k p_{\theta_k}(y)}\right) dy \\
&\quad = -\int_{\mathbb{R}^d} p_{\widehat{\theta}_1}(y) \log\left(\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right) dy \\
&\quad = -\mathbb{E}_{\widehat{\theta}_1}\left[\log\left\{\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\}\right] \\
&\quad \geq -\log \mathbb{E}_{\widehat{\theta}_1}\left\{\frac{\sum_{k=1}^K w_k p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\} && \text{(S17)} \\
&\quad = -\log\left[\sum_{k=1}^K w_k \mathbb{E}_{\widehat{\theta}_1}\left\{\frac{p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\}\right] \\
&\quad \geq -\log\left\{\sum_{k=1}^K w_k (1)\right\} && \text{(S18)} \\
&\quad = 0.
\end{aligned}
$$

(S17) holds by Jensen's inequality. (S18) holds by the following derivation:

$$
\begin{aligned}
&\mathbb{E}_{\widehat{\theta}_1}\left\{\frac{p_{\theta_k}(y)}{p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(y)}\right\} \\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2\right) \frac{\exp\left(-\frac{1}{2}\|y-\theta_k\|^2\right)}{\exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right)} dy \\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2 - \frac{1}{2}\|y-\widehat{\theta}_1+\widehat{\theta}_1-\theta_k\|^2 + \frac{1}{2}\|y-\widehat{\theta}_1+\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right) dy \\
&= \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\Bigg(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2 - (y-\widehat{\theta}_1)^T(\widehat{\theta}_1-\theta_k) - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 + (y-\widehat{\theta}_1)^T(\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|) + \\
&\qquad\qquad \frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\Bigg) dy \\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2\right) \int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|y-\widehat{\theta}_1\|^2 + (y-\widehat{\theta}_1)^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right) dy \\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2\right) \mathbb{E}_{\widehat{\theta}_1}\left[\exp\left\{(y-\widehat{\theta}_1)^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right\}\right] \\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 - \widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right) \mathbb{E}_{\widehat{\theta}_1}\left[\exp\left\{(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)^T y\right\}\right] \\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 - \widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|)\right) \exp\left\{\widehat{\theta}_1^T(\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|) + \frac{1}{2}\|\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right\} \\
&= \exp\left(\frac{1}{2}\|\widehat{\theta}_1-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2 - \frac{1}{2}\|\widehat{\theta}_1-\theta_k\|^2 + \frac{1}{2}\|\theta_k-\widehat{\theta}_1/\|\widehat{\theta}_1\|\|^2\right) \\
&= \exp\Bigg(\frac{1}{2}\|\widehat{\theta}_1\|^2 - \widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \frac{1}{2}\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2 - \frac{1}{2}\|\widehat{\theta}_1\|^2 + \widehat{\theta}_1^T\theta_k - \frac{1}{2}\|\theta_k\|^2 + \\
&\qquad \frac{1}{2}\|\theta_k\|^2 - \theta_k^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \frac{1}{2}\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2\Bigg) \\
&= \exp\left(\widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\|^2 - \widehat{\theta}_1^T\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k^T\widehat{\theta}_1/\|\widehat{\theta}_1\| + \widehat{\theta}_1^T\theta_k\right) \\
&= \exp\left\{(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k)\right\} \\
&\le \exp(0) \qquad\qquad \text{(S19)} \\
&= 1.
\end{aligned}
$$

To justify (S19), note that

$$(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k) = \left\|\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1\right\|\left\|\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k\right\|\cos(\gamma),$$

where γ is the angle between $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1$ and $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k$. Recall that Θ_0 is a spherical region, $\|\theta_k\| \in [0.5, 1]$, $\|\widehat{\theta}_1\| > 1$, and $\widehat{\theta}_1/\|\widehat{\theta}_1\|$ is on the outer border of Θ_0. Thus, γ will always be between 90° and 270°. (See Fig. S1.) This implies that $(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1)^T(\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k) \le 0$. □

Lemma 4. *Assume the doughnut null test setting. Let* $R_n = \prod_{Y_i \in \mathcal{D}_0}\{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i)\}$. *If* $\theta^* \in \Theta_0$, *then* $\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} \le \mathbb{1}(\|\overline{Y}_1\| > 1)$.

Proof. If $\mathcal{D}_1$ satisfies $\|\overline{Y}_1\| \le 1$, then

$$\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} = 0 = \mathbb{1}(\|\overline{Y}_1\| > 1).$$

Now suppose $\mathcal{D}_1$ satisfies $\|\overline{Y}_1\| > 1$. Then $\|\widehat{\theta}_1\| > 1$, and $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$ is the RIPR of $p_{\widehat{\theta}_1}$ onto the convex set of densities $\mathcal{P}_{\Theta_0}$, as proved in Lemma 3. Since $\theta^* \in \Theta_0$, $\widehat{\theta}_1 \in \Theta_1$, and $p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}$ is the RIPR of $p_{\widehat{\theta}_1}$

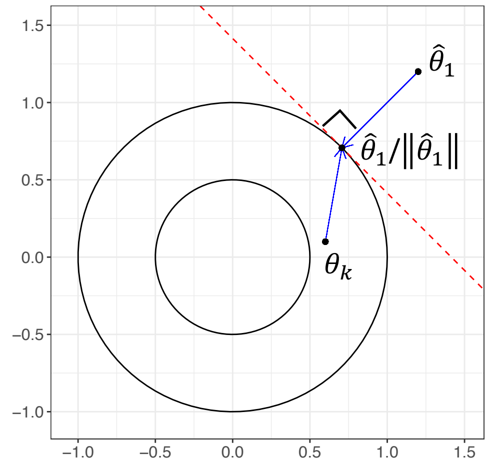

Fig. S1. Lemma 3 companion diagram. The angle between $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \widehat{\theta}_1$ and $\widehat{\theta}_1/\|\widehat{\theta}_1\| - \theta_k$ must be between $90°$ and $270°$.

onto $\mathcal{P}_{\Theta_0}$, we know $\mathbb{E}_{\theta^*}\{p_{\widehat{\theta}_1}(Y)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y)\} \leq 1$, as explained under *Approach 3: Subsampled hybrid LRT*. So

$$\begin{aligned}\mathbb{E}_{\theta^*}\{R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\} &= \mathbb{E}_{\theta^*}\left[\prod_{Y_i \in \mathcal{D}_0} \{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i)\}\right] \\ &\overset{iid}{=} \prod_{i=1}^{n/2} \mathbb{E}_{\theta^*}\left\{p_{\widehat{\theta}_1}(Y_i)/p_{\widehat{\theta}_1/\|\widehat{\theta}_1\|}(Y_i)\right\} \\ &\leq 1 \\ &= \mathbb{1}(\|\overline{Y}_1\| > 1).\end{aligned}$$

□

Theorem 6. *In the doughnut null hypothesis test setting, assume the subsampled test statistics* $U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \;/\; \mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{split})$ *and* $R_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_{1,b}) \;/\; \mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{RIPR})$*,* $1 \leq b \leq B$*. The test that rejects* H_0 *when*

$$\frac{1}{B}\sum_{b=1}^{B}\left\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\right\} \geq 1/\alpha$$

is a valid level α *test.*

Proof. Assume $\theta^* \in \Theta_0$. The probability of falsely rejecting H_0 is

$$
\begin{aligned}
&\mathbb{P}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\left\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\right\} \geq 1/\alpha\right] \\
&\leq \alpha\mathbb{E}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\left\{U_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\right\}\right] \\
&\leq \alpha\mathbb{E}_{\theta^*}\left[\frac{1}{B}\sum_{b=1}^{B}\left\{T_{n,b}(\theta^*)\mathbb{1}(\|\overline{Y}_{1,b}\| < 0.5) + \mathbb{1}(\|\overline{Y}_{1,b}\| \in [0.5, 1]) + R_{n,b}\mathbb{1}(\|\overline{Y}_{1,b}\| > 1)\right\}\right] \qquad \text{(S20)} \\
&= \alpha\mathbb{E}_{\theta^*}\left\{T_n(\theta^*)\mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n\mathbb{1}(\|\overline{Y}_1\| > 1)\right\} \\
&= \alpha\mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left\{T_n(\theta^*)\mathbb{1}(\|\overline{Y}_1\| < 0.5) \mid \mathcal{D}_1\right\}\right] + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5, 1]) + \alpha\mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left\{R_n\mathbb{1}(\|\overline{Y}_1\| > 1) \mid \mathcal{D}_1\right\}\right] \\
&\leq \alpha\mathbb{E}_{\theta^*}\left[\mathbb{1}(\|\overline{Y}_1\| < 0.5)\mathbb{E}_{\theta^*}\left\{T_n(\theta^*) \mid \mathcal{D}_1\right\}\right] + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5, 1]) + \alpha\mathbb{E}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| > 1)\} \qquad \text{(S21)} \\
&\leq \alpha\mathbb{E}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| < 0.5)\} + \alpha\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5, 1]) + \alpha\mathbb{P}_{\theta^*}\{\mathbb{1}(\|\overline{Y}_1\| > 1)\} \qquad \text{(S22)} \\
&= \alpha\left\{\mathbb{P}_{\theta^*}(\|\overline{Y}_1\| < 0.5) + \mathbb{P}_{\theta^*}(\|\overline{Y}_1\| \in [0.5, 1]) + \mathbb{P}_{\theta^*}(\|\overline{Y}_1\| > 1)\right\} \\
&= \alpha.
\end{aligned}
$$

(S20) holds because $\widehat{\theta}_{0,b}^{\text{split}} = \arg\max_{\theta \in \Theta_0} \mathcal{L}_{0,b}(\theta)$. Since $\theta^* \in \Theta_0$,

$$
U_{n,b} = \mathcal{L}_{0,b}(\widehat{\theta}_1)/\mathcal{L}_{0,b}(\widehat{\theta}_{0,b}^{\text{split}}) \leq \mathcal{L}_{0,b}(\widehat{\theta}_1)/\mathcal{L}_{0,b}(\theta^*) = T_{n,b}(\theta^*).
$$

(S21) holds by Lemma 4. (S22) holds because $\mathbb{E}_{\theta^*}\{T_n(\theta^*) \mid \mathcal{D}_1\} \leq 1$, as established by Theorem 1. □

S2. Derivations of Equations

Derivation of Equation 1. The usual likelihood ratio confidence set for $\theta^* \in \mathbb{R}^d$ is given by

$$
C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : 2\log\frac{\mathcal{L}(\overline{Y})}{\mathcal{L}(\theta)} \leq c_{\alpha,d}\right\},
$$

where $c_{\alpha,d}$ is the upper α quantile of the χ_d^2 distribution. $\overline{Y}$ is the sample mean of the Y_i observations, and it is also the MLE estimate for θ^*. We re-write this confidence set such that the squared radius of the set is apparent.

$$
\begin{aligned}
2\log\frac{\mathcal{L}(\overline{Y})}{\mathcal{L}(\theta)} &= 2\log\left(\frac{\Pi_{i=1}^{n}\exp\left(-\frac{1}{2}(Y_i-\overline{Y})^T(Y_i-\overline{Y})\right)}{\Pi_{i=1}^{n}\exp\left(-\frac{1}{2}(Y_i-\theta)^T(Y_i-\theta)\right)}\right)\\
&= 2\log\left(\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i-\overline{Y})^T(Y_i-\overline{Y})+\frac{1}{2}\sum_{i=1}^{n}(Y_i-\theta)^T(Y_i-\theta)\right)\right)\\
&= -\sum_{i=1}^{n}(Y_i-\overline{Y})^T(Y_i-\overline{Y})+\sum_{i=1}^{n}(Y_i-\theta)^T(Y_i-\theta)\\
&= \sum_{i=1}^{n}\left(-(Y_i-\overline{Y})^T(Y_i-\overline{Y})+(Y_i-\overline{Y}+\overline{Y}-\theta)^T(Y_i-\overline{Y}+\overline{Y}-\theta)\right)\\
&= \sum_{i=1}^{n}\Bigg(-(Y_i-\overline{Y})^T(Y_i-\overline{Y})+(Y_i-\overline{Y})^T(Y_i-\overline{Y})+\\
&\qquad 2(Y_i-\overline{Y})^T(\overline{Y}-\theta)+(\overline{Y}-\theta)^T(\overline{Y}-\theta)\Bigg)\\
&= n\|\overline{Y}-\theta\|^2.
\end{aligned}
$$

The final step holds because the first two terms cancel and the summation over the third term equals 0. Therefore,

$$
C_n^{\mathrm{LRT}}(\alpha)=\left\{\theta\in\Theta:\|\theta-\overline{Y}\|^2\le c_{\alpha,d}/n\right\}.
$$

□

Derivation of Equation 2. Let $\widehat{\theta}_1=\overline{Y}_1$ be the sample mean of the $n/2$ observations in $\mathcal{D}_1$. Where

$$
T_n(\theta)=\frac{\mathcal{L}_0(\widehat{\theta}_1)}{\mathcal{L}_0(\theta)},
$$

the universal confidence set using the split likelihood ratio statistic is

$$
C_n^{\mathrm{split}}(\alpha)=\left\{\theta\in\Theta:T_n(\theta)<\frac{1}{\alpha}\right\}.
$$

We also re-write this confidence set such that the squared radius of the set is apparent.

$$
\begin{aligned}
T_n(\theta) &= \frac{\Pi_{Y_i\in\mathcal{D}_0}\exp\left(-\frac{1}{2}(Y_{0i}-\widehat{\theta}_1)^T(Y_{0i}-\widehat{\theta}_1)\right)}{\Pi_{Y_i\in\mathcal{D}_0}\exp\left(-\frac{1}{2}(Y_{0i}-\theta)^T(Y_{0i}-\theta)\right)} \\
&= \exp\left(\sum_{Y_i\in\mathcal{D}_0}\left(-\frac{1}{2}(Y_{0i}-\overline{Y}_1)^T(Y_{0i}-\overline{Y}_1)+\frac{1}{2}(Y_{0i}-\theta)^T(Y_{0i}-\theta)\right)\right) \\
&= \exp\Bigg(\sum_{Y_i\in\mathcal{D}_0}\Bigg(-\frac{1}{2}(Y_{0i}-\overline{Y}_0+\overline{Y}_0-\overline{Y}_1)^T(Y_{0i}-\overline{Y}_0+\overline{Y}_0-\overline{Y}_1)+ \\
&\qquad\qquad \frac{1}{2}(Y_{0i}-\overline{Y}_0+\overline{Y}_0-\theta)^T(Y_{0i}-\overline{Y}_0+\overline{Y}_0-\theta)\Bigg)\Bigg) \\
&= \exp\Bigg(\sum_{Y_i\in\mathcal{D}_0}\Bigg(-\frac{1}{2}\left[(Y_{0i}-\overline{Y}_0)^T(Y_{0i}-\overline{Y}_0)+2(Y_{0i}-\overline{Y}_0)^T(\overline{Y}_0-\overline{Y}_1)+(\overline{Y}_0-\overline{Y}_1)^T(\overline{Y}_0-\overline{Y}_1)\right]+ \\
&\qquad\qquad \frac{1}{2}\left[(Y_{0i}-\overline{Y}_0)^T(Y_{0i}-\overline{Y}_0)+2(Y_{0i}-\overline{Y}_0)^T(\overline{Y}_0-\theta)+(\overline{Y}_0-\theta)^T(\overline{Y}_0-\theta)\right]\Bigg)\Bigg) \qquad \text{(S23)} \\
&= \exp\left(\sum_{Y_i\in\mathcal{D}_0}\left(-\frac{1}{2}(\overline{Y}_0-\overline{Y}_1)^T(\overline{Y}_0-\overline{Y}_1)+\frac{1}{2}(\overline{Y}_0-\theta)^T(\overline{Y}_0-\theta)\right)\right) \\
&= \exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2\right). \qquad \text{(S24)}
\end{aligned}
$$

The first and fourth terms of (S23) cancel, and the cross-product terms equal 0 upon taking the summation. (S24) holds because $\mathcal{D}_0$ contains $\frac{n}{2}$ elements. Therefore,

$$
\begin{aligned}
C_n^{\text{split}}(\alpha) &= \left\{\theta\in\Theta : T_n(\theta) < \frac{1}{\alpha}\right\} \\
&= \left\{\theta\in\Theta : \exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2\right) < \frac{1}{\alpha}\right\} \\
&= \left\{\theta\in\Theta : -\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\theta\|^2 < \log\left(\frac{1}{\alpha}\right)\right\} \\
&= \left\{\theta\in\Theta : \frac{n}{4}\|\overline{Y}_0-\theta\|^2 < \log\left(\frac{1}{\alpha}\right)+\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2\right\} \\
&= \left\{\theta\in\Theta : \|\overline{Y}_0-\theta\|^2 < \frac{4}{n}\log\left(\frac{1}{\alpha}\right)+\|\overline{Y}_0-\overline{Y}_1\|^2\right\}.
\end{aligned}
$$

□

Derivation of Equation 7. From the statement of $C_n^{\text{split}}(\alpha)$, we see that $r^2(C_n^{\text{split}}(\alpha)) = \frac{4}{n}\log(1/\alpha) + \|\overline{Y}_0-\overline{Y}_1\|^2$. Note that

$$
\|\overline{Y}_0-\overline{Y}_1\|^2 = \left\|\frac{2}{n}\sum_{i=1}^{n/2}(Y_{0i}-Y_{1i})\right\|^2 = \frac{4}{n}\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n/2}(Y_{0i}-Y_{1i})\right\|^2 \overset{d}{=} \frac{4}{n}\chi_d^2.
$$

To see why the last step holds, note that $Y_1,\ldots,Y_n \overset{iid}{\sim} N(\theta^*, I_d)$. So for any i, $Y_{0i}-Y_{1i} \overset{iid}{\sim} N(0, 2I_d)$. Then $\sum_{i=1}^{n/2}(Y_{0i}-Y_{1i}) \overset{iid}{\sim} N\left(0, \frac{n}{2}(2I_d)\right)$, and $\frac{1}{\sqrt{n}}\sum_{i=1}^{n/2}(Y_{0i}-Y_{1i}) \overset{iid}{\sim} N(0, I_d)$. This implies that $r^2(C_n^{\text{split}}(\alpha)) \overset{d}{=} \frac{4}{n}\log(1/\alpha) + \frac{4}{n}\chi_d^2$. Therefore, $\mathbb{E}[r^2(C_n^{\text{split}}(\alpha))] = \frac{4}{n}\log\left(\frac{1}{\alpha}\right) + \frac{4}{n}d$. □

Derivation of Equation 9. From equation 8, we know that

$$\mathbb{E}\left[\frac{r^2(C_n^{\text{split}}(\alpha))}{r^2(C_n^{\text{LRT}}(\alpha))}\right] = \frac{4\log(1/\alpha)+4d}{c_{\alpha,d}}.$$

For $d \geq 1$ and $\alpha \in (0,1)$, Inglot (2010) shows the upper bound

$$c_{\alpha,d} \leq d + 2\log\left(\frac{1}{\alpha}\right) + 2\sqrt{d\log\left(\frac{1}{\alpha}\right)}.$$

Also, for $d \geq 2$ and $\alpha \leq 0.17$, Inglot (2010) shows the lower bound

$$c_{\alpha,d} \geq d + 2\log\left(\frac{1}{\alpha}\right) - \frac{5}{2}.$$

Combining these facts, we see that for $d \geq 2$ and $\alpha \leq 0.17$,

$$\frac{4\log(1/\alpha)+4d}{2\log(1/\alpha)+d+2\sqrt{d\log(1/\alpha)}} \leq \mathbb{E}\left[\frac{r^2(C_n^{\text{split}}(\alpha))}{r^2(C_n^{\text{LRT}}(\alpha))}\right] \leq \frac{4\log(1/\alpha)+4d}{2\log(1/\alpha)+d-\frac{5}{2}}.$$

□

Derivation of Equation 10. From equation 8, we know that

$$\mathbb{E}\left[\frac{r^2(C_n^{\text{split}}(\alpha))}{r^2(C_n^{\text{LRT}}(\alpha))}\right] = \frac{4\log(1/\alpha)+4d}{c_{\alpha,d}}.$$

The lower bound of equation 10 is the same as the lower bound from equation 9. We consider the upper bound. Suppose $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$. Let $t = -2 + \sqrt{5 + 2\log(1/\alpha)}$. We will show that $c_{\alpha,1} \geq t^2$ in several steps:

Step 1: Show that $t^2 + 4t - 2 < 2\log(1/\alpha)$.

$$\begin{aligned}
t^2 + 4t - 2 &= \left(-2 + \sqrt{5+2\log(1/\alpha)}\right)^2 + 4(-2+\sqrt{5+2\log(1/\alpha)}) - 2 \\
&= 4 - 4\sqrt{5+2\log(1/\alpha)} + 5 + 2\log(1/\alpha) - 8 + 4\sqrt{5+2\log(1/\alpha)} - 2 \\
&= 2\log(1/\alpha) - 1 \\
&< 2\log(1/\alpha).
\end{aligned}$$

Step 2: Show that $\log(1/\alpha) > \frac{t^2}{2} + 2\log(t) + \log(\sqrt{2\pi})$. Starting with the result from Step 1,

$$\begin{aligned}
\log(1/\alpha) &> \frac{t^2}{2} + 2t - 1 \\
&\geq \frac{t^2}{2} + 2(\log(t)+1) - 1 \qquad \text{since } t \geq \log(t)+1 \text{ for } t > 0 \\
&= \frac{t^2}{2} + 2\log(t) + 1 \\
&> \frac{t^2}{2} + 2\log(t) + \log(\sqrt{2\pi}).
\end{aligned}$$

Step 3: Show that $t^2 - 1 \geq t$. We start by showing that $t \geq \frac{1}{2}(1+\sqrt{5})$ follows from our definitions of t and α:

$$
\begin{aligned}
&& \alpha &\leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right) \\
\iff && \frac{1}{\alpha} &\geq \exp\left(\frac{5(1+\sqrt{5})}{4}\right) \\
\iff && 8\log(1/\alpha) &\geq 10(1+\sqrt{5}) \\
\iff && 20 + 8\log(1/\alpha) &\geq 30 + 10\sqrt{5} \\
\iff && 4(5 + 2\log(1/\alpha)) &\geq 25 + 10\sqrt{5} + 5 \\
\iff && 2\sqrt{5 + 2\log(1/\alpha)} &\geq 5 + \sqrt{5} \\
\iff && -4 + 2\sqrt{5 + 2\log(1/\alpha)} &\geq 1 + \sqrt{5} \\
\iff && -2 + \sqrt{5 + 2\log(1/\alpha)} &\geq \frac{1}{2}(1+\sqrt{5}) \\
\iff && t &\geq \frac{1}{2}(1+\sqrt{5}).
\end{aligned}
$$

The roots of the convex function $t^2 - t - 1$ are at $t = (1 \pm \sqrt{5})/2$. At $t \geq (1/2)(1+\sqrt{5})$, we know $t^2 - 1 \geq t$.

Step 4: Show that $t^2 \leq c_{\alpha,1}$. Starting with the results of steps 2 and 3,

$$
\begin{aligned}
\log(t^2 - 1) - t^2/2 - \log(\sqrt{2\pi}) &> \log(t^2 - 1) + 2\log(t) + \log(\alpha) \\
&\geq 3\log(t) + \log(\alpha).
\end{aligned}
$$

Exponentiating,

$$
\left(t^2 - 1\right)\exp\left(-t^2/2\right)\left(\frac{1}{\sqrt{2\pi}}\right) \geq t^3\alpha.
$$

So

$$
\left(\frac{1}{t} - \frac{1}{t^3}\right)\exp\left(-t^2/2\right)\left(\frac{1}{\sqrt{2\pi}}\right) \geq \alpha.
$$

If $Z \sim N(0,1)$ and $X = Z^2 \sim \chi^2_1$, then using an inequality on $\mathbb{P}(Z \geq t)$ from Polland (2015),

$$
\mathbb{P}(X \geq t^2) = 2\mathbb{P}(Z \geq t) > \mathbb{P}(Z \geq t) \geq \left(\frac{1}{t} - \frac{1}{t^3}\right)\exp\left(-t^2/2\right)\left(\frac{1}{\sqrt{2\pi}}\right) \geq \alpha.
$$

This implies that $c_{\alpha,1} \geq t^2 = 2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}$. We conclude that for $d = 1$ and $\alpha \leq \exp\left(-\frac{5(1+\sqrt{5})}{4}\right)$,

$$
\frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + d + 2\sqrt{d\log(1/\alpha)}} \leq \mathbb{E}\left[\frac{r^2(C_n^{\text{split}}(\alpha))}{r^2(C_n^{\text{LRT}}(\alpha))}\right] \leq \frac{4\log(1/\alpha) + 4d}{2\log(1/\alpha) + 9 - 4\sqrt{5 + 2\log(1/\alpha)}}.
$$

□

Derivation of Equation 12. The usual LRT set is

$$C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 \leq c_{\alpha,d} \,/\, n\right\},$$

where $c_{\alpha,d}$ is the upper α quantile of the χ_d^2 distribution. Suppose we are testing $H_0 : \theta^* = 0$ versus $H_1 : \theta^* \neq 0$. The power of the usual LRT at the true θ^* is thus

$$\text{Power}(C_n^{\text{LRT}}(\alpha); \theta^*) = \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > c_{\alpha,d}/n\right).$$

We can express the power function of the usual LRT in terms of the CDF of a noncentral χ^2 distribution. Let us denote $\theta^* = (\theta_1^*, \ldots, \theta_d^*)$. We see that

$$n\|\overline{Y}\|^2 = \left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i\right\|^2 = \sum_{j=1}^d \left(\frac{1}{\sqrt{n}}\sum_{i=1}^n Y_{ij}\right)^2.$$

For each dimension j, $n^{-1/2}\sum_{i=1}^n Y_{ij} \sim N(\theta_j^*\sqrt{n}, 1)$. So this follows a non-central χ^2 distribution given by

$$n\|\overline{Y}\|^2 \stackrel{d}{=} \chi^2\left(df = d, \lambda = \sum_{j=1}^d n(\theta_j^*)^2\right) \stackrel{d}{=} \chi^2\left(df = d, \lambda = n\|\theta^*\|^2\right).$$

Let $\Phi(\cdot)$ represent the standard normal CDF. Suppose $X \sim \chi^2(df = d, \lambda = n\|\theta^*\|^2)$. As $d \to \infty$ or as $\lambda \to \infty$, it holds that

$$\frac{X - (d + n\|\theta^*\|^2)}{\sqrt{2(d + 2n\|\theta^*\|^2)}} \approx N(0, 1).$$

See Chun & Shapiro (2009). Using the Normal approximation to the non-central chi-squared CDF, the power of the usual LRT is

$$\begin{aligned}
\text{Power}(C_n^{\text{LRT}}(\alpha); \theta^*) &= \mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2 > c_{\alpha,d}/n\right) \\
&= \mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2 > c_{\alpha,d}\right) \\
&= \mathbb{P}_{\theta^*}\left(\frac{n\|\overline{Y}\|^2 - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}} > \frac{c_{\alpha,d} - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right) \\
&\approx 1 - \Phi\left(\frac{c_{\alpha,d} - d - n\|\theta^*\|^2}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right) \\
&= \Phi\left(\frac{d + n\|\theta^*\|^2 - c_{\alpha,d}}{\sqrt{2(d + 2n\|\theta^*\|^2)}}\right).
\end{aligned}$$

□

Derivation of Equation 13. Using methods from the derivation of equation 12, we can find a representation for the approximate power of the limiting subsampling LRT set as $B \to \infty$. From equation 4,

$$C_n^{\text{subsplit}}(\alpha) \approx \left\{\theta \in \Theta : \|\overline{Y} - \theta\|^2 < \frac{10}{3n}\log\left(\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right)\right\}$$

So the power of the limit of subsampling LRT is

$$\begin{aligned}
\text{Power}(C_n^{\text{subsplit}}(\alpha);\theta^*) &\approx \mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2 \ge \frac{10}{3}\log\left(\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right)\right)\\
&= \mathbb{P}_{\theta^*}\left(\frac{n\|\overline{Y}\|^2 - d - n\|\theta^*\|^2}{\sqrt{2(d+2n\|\theta^*\|^2)}} \ge \frac{(10/3)\log\left((5/2)^{d/2}(1/\alpha)\right) - d - n\|\theta^*\|^2}{\sqrt{2(d+2n\|\theta^*\|^2)}}\right)\\
&\approx \Phi\left(\frac{1}{\sqrt{2(d+2n\|\theta^*\|^2)}}\left(d + n\|\theta^*\|^2 - \frac{10}{3}\log\left(\left(\frac{5}{2}\right)^{d/2}\frac{1}{\alpha}\right)\right)\right).
\end{aligned}$$

□

S3. Simulated cross-fit sets with varying p_0

In the split LRT case, the optimal split proportion p_0^* (established in Theorem 3) converges to 0.5 as $d \to \infty$. This optimal split proportion minimizes the expected squared radius. Under general p_0, the cross-fit set is defined as

$$C_n^{\text{CF}}(\alpha) = \left\{\theta \in \Theta : \frac{1}{2}\left[\exp\left(-\frac{np_0}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{np_0}{2}\|\overline{Y}_0 - \theta\|^2\right) + \exp\left(-\frac{n(1-p_0)}{2}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n(1-p_0)}{2}\|\overline{Y}_1 - \theta\|^2\right)\right] < \frac{1}{\alpha}\right\}.$$

Noting the symmetry of the set $C_n^{\text{CF}}(\alpha)$, we conjecture that $p_0 = 0.5$ will minimize the expected squared diameter of the cross-fit set. Figure S2 presents examples of cross-fit sets at varying p_0 on a single sample of 1000 observations simulated from a $N(\vec{0}, I_2)$ distribution. We see that the regions with $p_0 \in \{0.5, 0.7\}$ have smaller diameters than the regions with $p_0 \in \{0.1, 0.3, 0.9\}$.

S4. Power of Tests of $H_0 : \|\theta^*\| \in [0.5, 1]$

S4.1. *Exact Formula for Power of Intersection Test*

In section 4, we present hypothesis tests for $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$. The power of the intersection method that we present is tractable. We derive a formula for the intersection method's power at θ^*. From the intersection method's description, we reject H_0 if and only if $C_n^{\text{LRT}}(\alpha) \cap (\mathcal{S}_1 \backslash \mathcal{S}_{0.5}) = \emptyset$, where $C_n^{\text{LRT}}(\alpha) = \left\{\theta \in \Theta : \|\theta - \overline{Y}\|^2 \le c_{\alpha,d}/n\right\}$. This is equivalent to rejecting H_0 if and only if $\widehat{\theta}^{\text{proj}} \notin C_n^{\text{LRT}}(\alpha)$, where

$$\widehat{\theta}^{\text{proj}} = \begin{cases} 0.5\,\overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| < 0.5 \\ \overline{Y} & : \|\overline{Y}\| \in [0.5, 1.0] \\ \overline{Y}/\|\overline{Y}\| & : \|\overline{Y}\| > 1 \end{cases}.$$

In **Case 2**, we have $\|\overline{Y}\| \in [0.5, 1]$. In this setting, it is always true that $\widehat{\theta}^{\text{proj}} = \overline{Y} \in C_n^{\text{LRT}}(\alpha)$. So we will never reject H_0 in this case. We consider **Case 1** ($\|\overline{Y}\| < 0.5$) and **Case 3** ($\|\overline{Y}\| > 1$). For $\|\theta^*\| \notin$

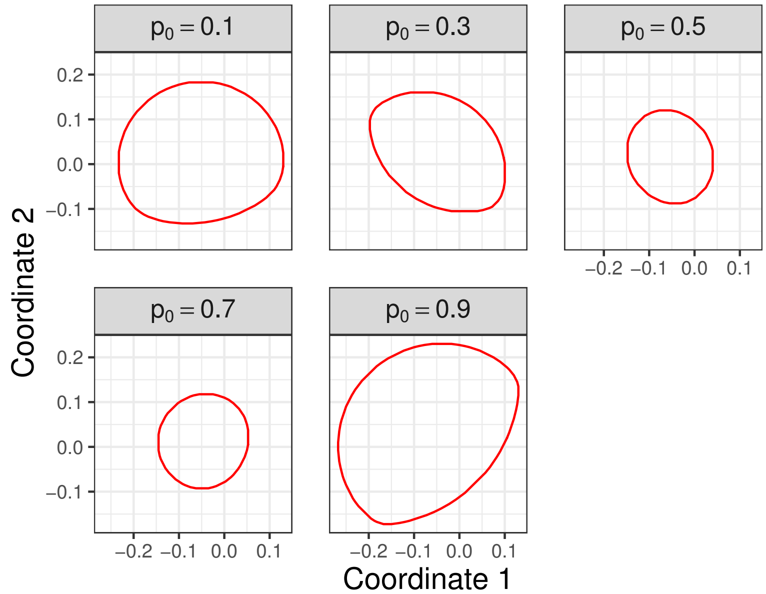

Fig. S2. Simulated cross-fit regions at varying p_0, using a single data sample.

$[0.5, 1.0]$, the power is given by

$$\text{Power}(\theta^*) = \mathbb{P}_{\theta^*}\left(\left\|\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| > 1\right) + \mathbb{P}_{\theta^*}\left(\left\|0.5\,\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 > c_{\alpha,d}/n,\ \|\overline{Y}\| < 0.5\right).$$

We know that $n\|\overline{Y}\|^2 \sim \chi^2(df = d, \lambda = n\|\theta^*\|^2)$. We will use this fact to write $\text{Power}(\theta^*)$ in terms of this non-central χ^2 CDF.

Case 1. Note that

$$\left\|0.5\,\overline{Y}/\|\overline{Y}\| - \overline{Y}\right\|^2 = \frac{\overline{Y}^T\overline{Y}}{4\|\overline{Y}\|^2} - 2\frac{\overline{Y}^T\overline{Y}}{2\|\overline{Y}\|} + \|\overline{Y}\|^2 = \frac{1}{4} - \|\overline{Y}\| + \|\overline{Y}\|^2 = \left(\|\overline{Y}\| - \frac{1}{2}\right)^2.$$

Then we write

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left\|0.5\,\overline{Y}/\|\overline{Y}\|-\overline{Y}\right\|^2>c_{\alpha,d}/n,\ \|\overline{Y}\|<1/2\right)\\
&=\mathbb{P}_{\theta^*}\left(\left(\|\overline{Y}\|-1/2\right)^2>c_{\alpha,d}/n,\ \|\overline{Y}\|<1/2\right)\\
&=\mathbb{P}_{\theta^*}\left(1/2-\|\overline{Y}\|>(c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|<1/2\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|<1/2-(c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|<1/2\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|<1/2-(c_{\alpha,d}/n)^{1/2}\right)\\
&=\mathbb{1}\left(c_{\alpha,d}/n<1/4\right)\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|<1/2-(c_{\alpha,d}/n)^{1/2}\right)\\
&=\mathbb{1}\left(n>4c_{\alpha,d}\right)\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2<1/4-\sqrt{c_{\alpha,d}/n}+c_{\alpha,d}/n\right)\\
&=\mathbb{1}\left(n>4c_{\alpha,d}\right)\mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2<n/4-\sqrt{nc_{\alpha,d}}+c_{\alpha,d}\right)\\
&=\mathbb{1}\left(n>4c_{\alpha,d}\right)F_{d,n\|\theta^*\|^2}\left(n/4-\sqrt{nc_{\alpha,d}}+c_{\alpha,d}\right), \qquad \text{(S25)}
\end{aligned}$$

where $F_{d,n\|\theta^*\|^2}$ is the non-central $\chi^2(df=d,\lambda=n\|\theta^*\|^2)$ CDF.

Case 3. Note that

$$\left\|\overline{Y}/\|\overline{Y}\|-\overline{Y}\right\|^2=\frac{\overline{Y}^T\overline{Y}}{\|\overline{Y}\|^2}-2\frac{\overline{Y}^T\overline{Y}}{\|\overline{Y}\|}+\|\overline{Y}\|^2=1-2\|\overline{Y}\|+\|\overline{Y}\|^2=\left(\|\overline{Y}\|-1\right)^2.$$

Then we write

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\left\|\overline{Y}/\|\overline{Y}\|-\overline{Y}\right\|^2>c_{\alpha,d}/n,\ \|\overline{Y}\|>1\right)\\
&=\mathbb{P}_{\theta^*}\left(\left(\|\overline{Y}\|-1\right)^2>c_{\alpha,d}/n,\ \|\overline{Y}\|^2>1\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|-1>(c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|^2>1\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|>1+(c_{\alpha,d}/n)^{1/2},\ \|\overline{Y}\|^2>1\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2>1+(2/\sqrt{n})c_{\alpha,d}^{1/2}+c_{\alpha,d}/n,\ \|\overline{Y}\|^2>1\right)\\
&=\mathbb{P}_{\theta^*}\left(\|\overline{Y}\|^2>1+(2/\sqrt{n})c_{\alpha,d}^{1/2}+c_{\alpha,d}/n\right)\\
&=\mathbb{P}_{\theta^*}\left(n\|\overline{Y}\|^2>n+2\sqrt{nc_{\alpha,d}}+c_{\alpha,d}\right)\\
&=1-F_{d,n\|\theta^*\|^2}(n+2\sqrt{nc_{\alpha,d}}+c_{\alpha,d}), \qquad \text{(S26)}
\end{aligned}$$

where $F_{d,n\|\theta^*\|^2}$ is the non-central $\chi^2(df=d,\lambda=n\|\theta^*\|^2)$ CDF.

For a given $\|\theta^*\|\notin[0.5,1]$, our calculation of Power(θ^*) is given by (S26) + (S25). That is,

$$\begin{aligned}
\text{Power}(\theta^*)=1-F_{d,n\|\theta^*\|^2}(n+2\sqrt{nc_{\alpha,d}}+c_{\alpha,d})+&\\
\mathbb{1}\left(n>4c_{\alpha,d}\right)F_{d,n\|\theta^*\|^2}&\left(n/4-\sqrt{nc_{\alpha,d}}+c_{\alpha,d}\right).
\end{aligned}$$

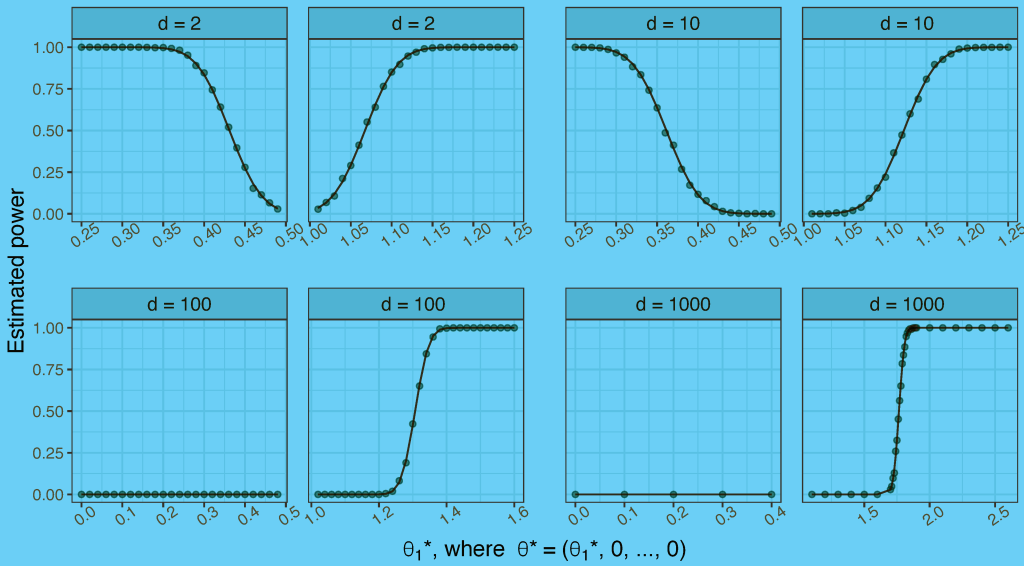

Fig. S3. Calculated power of $H_0 : \|\theta^*\| \in [0.5, 1.0]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1.0]$ using the intersection method. We compare the simulated power to the calculation (S26) + (S25). The points correspond to the simulated power, and the curves trace out the calculated power.

Figure S3 compares this calculated power to the simulated power of the intersection method from Figure 7. The points correspond to the simulated power, and the curves trace out the calculated power. The calculated and simulated powers align.

S4.2. *Cases of the Subsampled Hybrid LRT*

The subsampled hybrid test of $H_0 : \|\theta^*\| \in [0.5, 1]$ versus $H_1 : \|\theta^*\| \notin [0.5, 1]$ takes one of three approaches within each repeated subsample:

1. If $\|\overline{Y}_{1,b}\| < 0.5$, use the split LRT statistic U_n on the b^{th} subsample.
2. If $\|\overline{Y}_{1,b}\| \in [0.5, 1]$, set the b^{th} subsample's test statistic to 1.
3. If $\|\overline{Y}_{1,b}\| > 1$, use the RIPR LRT statistic R_n on the b^{th} subsample.

Figure S4 shows the proportion of these three cases that make up the hybrid test. We consider all $\|\theta^*\|$ values from Fig. 7 of the main paper, as well as cases where $\|\theta^*\|$ is within the null region. At any given value of d and $\|\theta^*\|$, the three proportions sum to 1. Interestingly, although $\|\overline{Y}_{1,b}\| < 0.5$ approximately 95% of the time when $\|\theta^*\| = 0$ and $d = 100$, the hybrid test has approximately zero power at that choice of parameters. We derive this fact in section S4.3. In addition, when $d = 1000$ we see that $\|\overline{Y}_{1,b}\| > 1$ in all simulations, even at $\theta^* = 0$. In section S4.4, we see why this setting has approximately zero power as well.

S4.3. *Hybrid power when $\theta^* = 0$, $d = 100$, and $n = 1000$*

When $\theta^* = 0$, $d = 100$, and $n = 1000$, Fig. S4 shows that $\|\overline{Y}_{1,b}\| < 0.5$ (case 1) occurs with probability of approximately 0.95, and $\|\overline{Y}_{1,b}\| \in [0.5, 1]$ (case 2) occurs with probability of approximately 0.05. At these parameters, the hybrid method has power of approximately 0, as shown in Fig. 7 in the main paper. We consider the power of the hybrid method at a single split of the data:

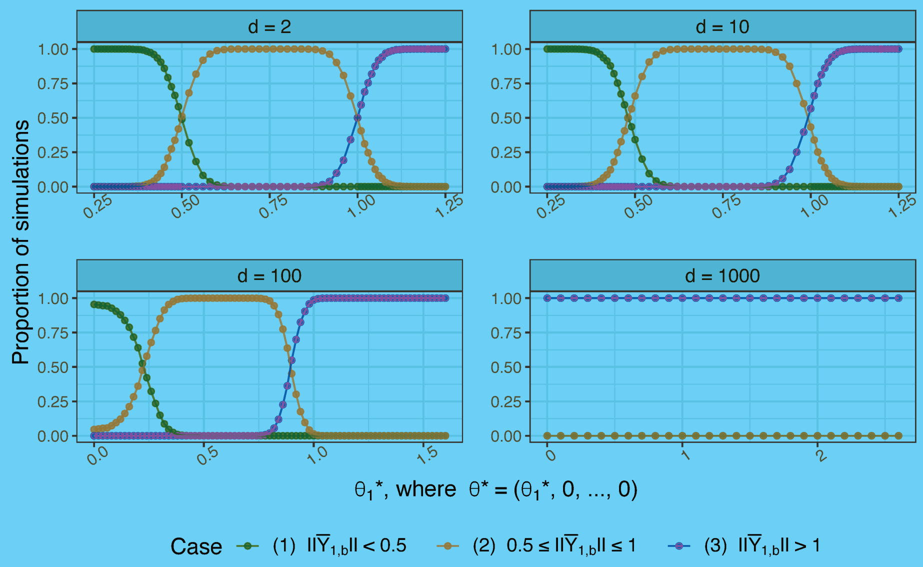

Fig. S4. Proportions of three cases that compose the hybrid LRT. We set $\alpha = 0.10$ and $n = 1000$, and we perform 1000 simulations at each value of $\|\theta^*\|$. We subsample $B = 100$ times.

$$
\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \\
&= \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| < 0.5)}_{A_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - 0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2\right) \geq \frac{1}{\alpha} \,\Big|\, \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| < 0.5\right)}_{A_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| \in [0.5, 1])}_{B_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_0\|^2\right) \geq \frac{1}{\alpha} \,\Big|\, \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| \in [0.5, 1]\right)}_{B_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| > 1)}_{C_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_0/\|\overline{Y}_0\|\|^2\right) \geq \frac{1}{\alpha} \,\Big|\, \|\overline{Y}_1\| < 0.5, \|\overline{Y}_0\| > 1\right)}_{C_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| \in [0.5, 1])}_{D_1} \underbrace{\mathbb{P}_{\theta^*=0}(1 \geq 1/\alpha \mid \|\overline{Y}_1\| \in [0.5, 1])}_{D_2} + \\
&\quad \underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| > 1)}_{E_1} \underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1/\|\overline{Y}_1\|\|^2\right) \geq \frac{1}{\alpha} \,\Big|\, \|\overline{Y}_1\| > 1\right)}_{E_2}.
\end{aligned}
$$

The probabilities B_2 and D_2 equal 0. In addition,

$$
\begin{aligned}
\mathbb{P}_{\theta^*=0}(\|\overline{Y}_0\| > 1) &= \mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\| > 1) \\
&= \mathbb{P}_{\theta^*=0}\left(\frac{n}{2}\|\overline{Y}_1\|^2 > \frac{n}{2}\right) \\
&= \mathbb{P}(\chi^2_{df=100} > 1000/2) \\
&\approx 0.
\end{aligned}
$$

So C_1 and E_1 are also approximately 0. That means we only need to consider $A_1 A_2$. It will be easier to work with the product of these two probabilities:

$$
\begin{aligned}
A_1A_2 &= \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2\right)\geq\frac{1}{\alpha},\ \|\overline{Y}_1\|<0.5,\ \|\overline{Y}_0\|<0.5\right)\\
&\leq \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0-\overline{Y}_1\|^2<\|\overline{Y}_0-0.5\overline{Y}_0/\|\overline{Y}_0\|\|^2, \|\overline{Y}_0\|<0.5\right)\\
&\leq \mathbb{P}_{\theta^*=0}(\|\overline{Y}_0-\overline{Y}_1\|^2<0.25)\\
&= \mathbb{P}((4/n)\chi^2_{df=100}<1/4)\\
&= \mathbb{P}(\chi^2_{df=100}<1000/16)\\
&\approx 0.001.
\end{aligned}
$$

This means that at a single split of the data, the power at $\|\theta^*\|=0$, $d=100$, and $n=1000$ is

$$
\mathbb{P}_{\theta^*=0}(U_n\mathbb{1}(\|\overline{Y}_1\|<0.5)+\mathbb{1}(\|\overline{Y}_1\|\in[0.5,1])+R_n\mathbb{1}(\|\overline{Y}_1\|>1)\geq 1/\alpha)\leq 0.001.
$$

S4.4. *Hybrid power when $\theta^*=0$, $d=1000$, and $n=1000$*

When $\theta^*=0$, $d=1000$, and $n=1000$, we see that the hybrid method selects case 3 ($\|\overline{Y}_{1,b}\|>1$) in all simulations. This is essentially choosing the wrong case, since $\|\theta^*\|=0<0.5$. Numerically, we can show that the hybrid method will have power of approximately 0 at these parameters. Again, we consider a single split of the data.

$$
\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n\mathbb{1}(\|\overline{Y}_1\|<0.5)+\mathbb{1}(\|\overline{Y}_1\|\in[0.5,1])+R_n\mathbb{1}(\|\overline{Y}_1\|>1)\geq 1/\alpha)\\
&\quad=\underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\|<0.5)}_{A_1}\underbrace{\mathbb{P}_{\theta^*=0}\left(U_n\geq 1/\alpha\,\middle|\,\|\overline{Y}_1\|<0.5\right)}_{A_2}+\\
&\qquad\underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\|\in[0.5,1])}_{B_1}\underbrace{\mathbb{P}_{\theta^*=0}(1\geq 1/\alpha\mid\|\overline{Y}_1\|\in[0.5,1])}_{B_2}+\\
&\qquad\underbrace{\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\|>1)}_{C_1}\underbrace{\mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1/\|\overline{Y}_1\|\|^2\right)\geq\frac{1}{\alpha}\,\middle|\,\|\overline{Y}_1\|>1\right)}_{C_2}.
\end{aligned}
$$

The probability B_2 equals 0. In addition, A_1 is approximately 0 because

$$
\begin{aligned}
\mathbb{P}_{\theta^*=0}(\|\overline{Y}_1\|<0.5) &= \mathbb{P}_{\theta^*=0}\left((n/2)\|\overline{Y}_1\|^2<n/8\right)\\
&= \mathbb{P}(\chi^2_{df=1000}<1000/8)\\
&\approx 0.
\end{aligned}
$$

So the probability of rejecting H_0 at this choice of parameters is approximately

$$\begin{aligned}
C_1C_2 &= \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1\|^2+\frac{n}{4}\|\overline{Y}_0-\overline{Y}_1/\|\overline{Y}_1\|\|^2\right)\geq\frac{1}{\alpha},\|\overline{Y}_1\|>1\right)\\
&\leq \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0-\overline{Y}_1\|^2<\|\overline{Y}_0-\overline{Y}_1/\|\overline{Y}_1\|\|^2,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(\|\overline{Y}_0\|^2-2\overline{Y}_0^T\overline{Y}_1+\|\overline{Y}_1\|^2<\|\overline{Y}_0\|^2-2\overline{Y}_0^T\overline{Y}_1/\|\overline{Y}_1\|+1,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1/\|\overline{Y}_1\|-1)+\|\overline{Y}_1\|^2<1,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1-\|\overline{Y}_1\|)/\|\overline{Y}_1\|<1-\|\overline{Y}_1\|^2,\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1(1-\|\overline{Y}_1\|)/\|\overline{Y}_1\|<(1-\|\overline{Y}_1\|)(1+\|\overline{Y}_1\|),\|\overline{Y}_1\|>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(2\overline{Y}_0^T\overline{Y}_1>\|\overline{Y}_1\|(1+\|\overline{Y}_1\|),\|\overline{Y}_1\|>1\right)\\
&\leq \mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T\overline{Y}_1>1\right).
\end{aligned}$$

Let $\sigma=1/\sqrt{500}$. Since $\overline{Y}_0$ and $\overline{Y}_1$ are averages of 500 $N(0,I_d)$ random variables, we see that $\overline{Y}_0\sim N(0,\sigma^2I_d)$ and $\overline{Y}_1\sim N(0,\sigma^2I_d)$. Let $\lambda=-d/2+(1/2)\sqrt{d^2+4/\sigma^4}$. (This choice of λ minimizes $\mathbb{E}[\exp(\lambda\overline{Y}_0^T\overline{Y}_1)]/\exp(\lambda)$ out of $\lambda>0$.) Let $\nu=\sigma/(1-\sigma^4\lambda^2)^{1/2}$. We derive

$$\begin{aligned}
&\mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T\overline{Y}_1>1\right)\\
&= \mathbb{P}_{\theta^*=0}\left(\exp\left(\lambda\overline{Y}_0^T\overline{Y}_1\right)>\exp(\lambda)\right)\\
&\leq \mathbb{E}_{\theta^*=0}\left[\exp\left(\lambda\overline{Y}_0^T\overline{Y}_1\right)\right]/\exp(\lambda)\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^d|\sigma^2I_d|}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_0\|^2-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2+\lambda\overline{Y}_0^T\overline{Y}_1\right)d\overline{Y}_0d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\left\{\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_0\|^2+\lambda\overline{Y}_0^T\overline{Y}_1\right)d\overline{Y}_0\right\}d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\left\{\mathbb{E}\left[\exp((\lambda\overline{Y}_1)^T\overline{Y}_0)\mid\overline{Y}_1\right]\right\}d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2\sigma^2}\|\overline{Y}_1\|^2\right)\exp\left(\frac{1}{2}\lambda^2\sigma^2\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2}\left(\frac{1}{\sigma^2}-\sigma^2\lambda^2\right)\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2}\left(\frac{1-\sigma^4\lambda^2}{\sigma^2}\right)\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&= \exp(-\lambda)\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\sigma^2I_d|^{1/2}}\exp\left(-\frac{1}{2\nu^2}\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&= \exp(-\lambda)\frac{|\nu^2I_d|^{1/2}}{|\sigma^2I_d|^{1/2}}\int_{\mathbb{R}^d}\frac{1}{(2\pi)^{d/2}|\nu^2I_d|^{1/2}}\exp\left(-\frac{1}{2\nu^2}\|\overline{Y}_1\|^2\right)d\overline{Y}_1\\
&= \exp(-\lambda)(\nu/\sigma)^d\\
&\approx \exp(-207)(1.1)^{1000}\\
&\approx 0.
\end{aligned}$$

At a single split of the data, the power at $\|\theta^*\| = 0$, $d = 1000$, and $n = 1000$ is approximately 0 because

$$\begin{aligned}
&\mathbb{P}_{\theta^*=0}(U_n \mathbb{1}(\|\overline{Y}_1\| < 0.5) + \mathbb{1}(\|\overline{Y}_1\| \in [0.5, 1]) + R_n \mathbb{1}(\|\overline{Y}_1\| > 1) \geq 1/\alpha) \\
&\quad \approx \mathbb{P}_{\theta^*=0}\left(\exp\left(-\frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1\|^2 + \frac{n}{4}\|\overline{Y}_0 - \overline{Y}_1/\|\overline{Y}_1\|\|^2\right) \geq \frac{1}{\alpha}, \|\overline{Y}_1\| > 1\right) \\
&\quad \leq \mathbb{P}_{\theta^*=0}\left(\overline{Y}_0^T \overline{Y}_1 > 1\right) \\
&\quad \approx 0.
\end{aligned}$$