Optimal Detection of Sparse Mixtures Against a Given Null Distribution

Tony T. Cai and Yihong Wu, *Member, IEEE*

***Abstract*— Detection of sparse signals arises in a wide range of modern scientific studies. The focus so far has been mainly on Gaussian mixture models. In this paper, we consider the detection problem under a general sparse mixture model and obtain explicit expressions for the detection boundary under mild regularity conditions. In addition, for Gaussian null hypothesis, we establish the adaptive optimality of the higher criticism procedure for all sparse mixtures satisfying the same conditions. In particular, the general results obtained in this paper recover and extend in a unified manner the previously known results on sparse detection far beyond the conventional Gaussian model and other exponential families.**

***Index Terms*— Hypothesis testing, high-dimensional statistics, sparse mixture, higher criticism, adaptive tests, total variation, Hellinger distance.**

I. Introduction

DETECTION of sparse mixtures is an important problem that arises in many scientific applications such as signal processing [1], biostatistics [2], and astrophysics [3], [4], where the goal is to determine the existence of a signal which only appears in a small fraction of the noisy data. For example, topological defects and Doppler effects manifest themselves as non-Gaussian convolution component in the Cosmic Microwave Background (CMB) temperature fluctuations. Detection of non-Gaussian signatures are important to identify cosmological origins of many phenomena [4]. Another example is disease surveillance where it is critical to discover an outbreak when the infected population is small [5]. The detection problem is of significant interest also because it is closely connected to a number of other important problems including estimation, screening, large-scale multiple testing, and classification. See, for example, [2], and [6]–[9].

A. Detection of Sparse Binary Vectors

One of the earliest work on sparse mixture detection dates back to Dobrushin [1], who considered the following problem originating from multi-channel detection in radiolocation. Let $\mathrm{Ray}(\alpha)$ denote the Rayleigh distribution with the density $\frac{2y}{\alpha}\exp(-\frac{y^2}{\alpha})$, $y \geq 0$. Let $\{Y_i\}_{i=1}^n$ be independently distributed according to $\mathrm{Ray}(\alpha_i)$, representing the random voltages observed on the n channels. In the absence of noise, α_i's are all equal to one, the nominal value; while in the presence of signal, exactly one of the α_i's becomes a known value $\alpha > 1$. Denoting the uniform distribution on $[n]$ by U_n, the goal is to test the following competing hypotheses:

$$\begin{aligned} H_0^{(n)} &: \alpha_i = 1, i \in [n], \\ \text{v.s. } H_1^{(n)} &: \alpha_i = 1 + (\alpha - 1)\mathbf{1}_{\{i=J\}}, J \sim U_n. \end{aligned} \tag{1}$$

Since the signal only appears once out of the n samples, in order for the signal to be distinguishable from noise, it is necessary for the amplitude α to grow with the sample size n (in fact, at least logarithmically). By proving that the log-likelihood ratio converges to a stable distribution in the large-n limit, Dobrushin [1] obtained sharp asymptotics of the smallest α in order to achieve the desired false alarm and miss detection probabilities. Similar results are obtained in the continuous-time Gaussian setting by Burnashev and Begmatov [10].

Subsequent important work include Ingster [11] and Donoho and Jin [8], which focused on detecting a sparse binary vector in the presence of Gaussian observation noise. The problem can be formulated as follows. Given a random sample $\{Y_1, ..., Y_n\}$, one wishes to test the hypotheses

$$\begin{aligned} H_0^{(n)} &: Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), i \in [n], \\ \text{v.s. } H_1^{(n)} &: Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)\mathcal{N}(0,1) + \epsilon_n \mathcal{N}(\mu_n, 1), \quad i \in [n] \end{aligned} \tag{2}$$

where the non-null proportion ϵ_n is calibrated according to

$$\epsilon_n = n^{-\beta}, \quad \frac{1}{2} < \beta < 1, \tag{3}$$

and the non-null effect μ_n grows with the sample size according to

$$\mu_n = \sqrt{2r \log n}, \quad r > 0. \tag{4}$$

Equivalently, one can write

$$Y_i = X_i + Z_i \tag{5}$$

where $Z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ is the observation noise. Under the null hypothesis, the mean vector $X^n = (X_1, \ldots, X_n)$ is equal to zero; under the alternative, X^n is a non-zero sparse binary vector with $X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)\delta_0 + \epsilon_n \delta_{\mu_n}$, where δ_a denotes the

Manuscript received November 9, 2012; revised January 12, 2014; accepted January 17, 2014. Date of publication February 3, 2014; date of current version March 13, 2014. This work was supported in part by NSF FRG under Grant DMS-0854973, in part by NSF under Grant DMS-1208982, and in part by NIH under Grant R01 CA127334.

T. T. Cai is with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA (e-mail: tcai@wharton.upenn.edu).

Y. Wu was with the Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104 USA. He is now with the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: yihongwu@illinois.edu).

Communicated by G. Moustakides, Associate Editor for Detection and Estimation.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIT.2014.2304295

point mass at a. This formulation is closely related to the following minimax composite testing problem:

$$\begin{aligned} &H_0 : Y^n \sim \mathcal{N}(0, \mathbf{I}_n), \\ \text{v.s. } &H_1 : Y^n \sim \mathcal{N}(\theta, \mathbf{I}_n), \theta \in \{0, \mu_n\}^n, \theta \text{ is } n\epsilon_n\text{-sparse}, \end{aligned} \tag{6}$$

where the worst-case prior for θ is an n-fold product of Bernoulli distributions.

The main objectives of studying the sparse detection problem (2) are two-fold.

1) Determine the *detection boundary*, which gives the smallest possible signal strength r as a function of the sparsity parameter β, denoted by denoted by $r^*(\beta)$, such that reliable detection is possible, in the sense that the sum of Type-I and II error probabilities vanishes as $n \to \infty$. The strict epigraph $\{(\beta, r) : r > r^*(\beta)\}$ is known as the *detectable region*.
2) Construct *adaptive* optimal tests, which achieve vanishing probability of error for all values of (r, β) inside the detectable region. Although the Neyman-Pearson lemma asserts that the optimal test is a likelihood ratio test, evaluating the likelihood ratio involves parameters in the alternative distribution, which might not be available to the statistician in practice. Therefore it is highly desirable to construct *adaptive* testing procedures which are simultaneously optimal for a class of alternatives. This problem is also known as *universal hypothesis testing* (c.f. [12]–[14] and the references therein, for results on discrete alphabets).

The detection boundary of (2) is given by the following function [8], [11]:

$$r^*(\beta) = \begin{cases} \beta - \frac{1}{2} & \frac{1}{2} < \beta \le \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2 & \frac{3}{4} < \beta < 1. \end{cases} \tag{7}$$

Therefore, the hypotheses in (2) can be tested with vanishing probability of error if $r > r^*(\beta)$; Conversely, if $r < r^*(\beta)$, the hypotheses are asymptotically indistinguishable and the sum of Type-I and II error probabilities goes to one for all tests. Furthermore, the fraction of the non-zero mean is so small that most tests based on empirical moments have no power in detection. To construct adaptive optimal procedures, Ingster [15], [16] considered generalized likelihood ratio tests over a growing discretized set of (β, r)-pairs and established its asymptotic adaptive optimality. A more elegant solution is provided by Donoho and Jin [8], who proposed an adaptive testing procedure based on Tukey's higher criticism statistic and showed that it attains the optimal detection boundary (7) without requiring the knowledge of the unknown parameters (β, r).

The above results have been generalized along various directions within the framework of two-component Gaussian mixtures. Jager and Wellner [17] proposed a family of goodness-of-fit tests based on the Rényi divergences [18, p. 554], including the higher criticism test as a special case, which achieve the optimal detection boundary adaptively. The detection boundary with correlated noise with known covariance matrices was established in [19] which also proposed a modified version of the higher criticism that achieves the corresponding optimal boundary. In a related setup, [20]–[22] considered detecting a signal with a known geometric shape in Gaussian noise. Minimax estimation of the non-null proportion ϵ_n was studied in Cai, Jin and Low [7].

The setup of [8] and [11] specifically focuses on the two-point Gaussian mixtures. Although [8] and [11] provide insightful results for sparse signal detection, the setting is highly restrictive and idealized. In particular, a major limitation that the signal strength must be constant under the alternative, i.e., the mean vector X^n takes constant value μ_n on its support. In many applications, the signal itself varies among the non-null portion of the samples. A natural question is the following: What is the detection boundary if μ_n varies under the alternative, say with a distribution P_n? Motivated by these considerations, the following heteroscedastic Gaussian mixture model was considered in Cai, Jeng and Jin [6]:

$$\begin{aligned} &H_0^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1) \\ \text{v.s. } &H_1^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)\mathcal{N}(0, 1) + \epsilon_n \mathcal{N}(\mu_n, \sigma^2). \end{aligned} \tag{8}$$

In this case, [6, Theorems 2.1 and 2.2] showed that reliable detection is possible if and only if $r > r^*(\beta, \sigma^2)$ where $r^*(\beta, \sigma^2)$ is given by

$$r^*(\beta, \sigma^2) = \begin{cases} (2-\sigma^2)(\beta - \frac{1}{2}) & \frac{1}{2} < \beta \le 1 - \frac{\sigma^2}{4}, \sigma^2 < 2 \\ (1 - \sigma\sqrt{1-\beta})_+^2 & \text{otherwise.} \end{cases} \tag{9}$$

where $x_+ \triangleq \max(x, 0)$. It was also shown that the optimal detection boundary can be achieved by a double-sided version of the higher criticism test.

B. Detection of General Sparse Mixture

Although the setup in Cai, Jeng and Jin [6] is more general than that considered in [11] and [8], it is still restricted to the two-component Gaussian mixtures. In many applications such as the aforementioned multi-channel detection [1] and astrophysical problems [4], the sparse signal may not be binary and the distribution may not be Gaussian. In the present paper, we consider the problem of sparse mixture detection in a general framework where the distributions are not necessarily Gaussian and the non-null effects are not necessarily a binary vector. More specifically, given a random sample $Y^n = (Y_1, ..., Y_n)$, we wish to test the following hypotheses

$$\begin{aligned} &H_0^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} Q_n \\ \text{v.s. } &H_1^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)Q_n + \epsilon_n G_n \end{aligned} \tag{10}$$

where Q_n is the null distribution and G_n is a distribution modeling the statistical variations of the non-null effects. The non-null proportion $\epsilon_n \in (0, 1)$ is calibrated according to (3).

Following the setup in [8] and [11], our goal is to determine when the sparse mixture in (10) is *detectable*, in the sense that the sum of Type-I and II error probabilities vanishes for some tests. It should be noted that there are various other formulations for optimal hypotheses testing, e.g., the Neyman-Pearson criterion which minimizes the Type-II error under fixed Type-I error probability, and the Chernoff formulation dealing with the optimal tradeoff between the speed of decay of Type-I and II error probabilities. For simplicity as well as

consistency with previous work, in this paper we focus on the sum of Type-I and II error and determine the corresponding detection boundary.

The main contribution of the present paper is as follows: First we obtain an explicit formula for the fundamental limit of the general testing problem (10) under mild technical conditions on the mixture, which are in particular satisfied by the Gaussian and generalized Gaussian null distributions. Rather than obtaining the limiting distribution of the log-likelihood ratio near the detection boundary as in [8], [23] which relies on the normality assumption of the null distribution, our analysis is based on analyzing the sharp asymptotics of the Hellinger distance between the null and the mixed alternative. Furthermore, under the Gaussian null, we also establish the adaptive optimality of the higher criticism procedure for all sparse mixtures satisfying the same regularity conditions. In particular, the general results obtained in this paper recover and extend all previously mentioned detection boundary results in a unified manner. The results also generalize the optimal adaptivity of the higher criticism procedure far beyond the original equal-signal-strength Gaussian setup in [8], [11] and the heteroscedastic extension in [6]. In the most general case, it turns out that detection boundary is determined by the asymptotic behavior of the log-likelihood ratio $\log \frac{\mathrm{d}G_n}{\mathrm{d}Q_n}$ evaluated at an appropriate quantile of the null distribution Q_n.

Although our general approach does not rely on the Gaussianity of the model, it is however instructive to begin by considering the special case of *sparse normal mixture* with $Q_n = \mathcal{N}(0,1)$, i.e.,

$$\begin{aligned} H_0^{(n)} &: Y_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\ \text{v.s. } H_1^{(n)} &: Y_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)\mathcal{N}(0,1) + \epsilon_n G_n \end{aligned} \tag{11}$$

The special case of the *convolution model* is interesting since it corresponds to detecting a sparse signal from additive Gaussian noise, where

$$G_n = P_n * \mathcal{N}(0,1) \tag{12}$$

is a standard normal mixture and $*$ denotes the convolution of two distributions. Indeed in this case the hypotheses (11) can be equivalently expressed via the additive-noise model (5), where $X_i = 0$ under the null and $X_i \overset{\text{i.i.d.}}{\sim} (1-\epsilon_n)\delta_0 + \epsilon_n P_n$ under the alternative. Based on the noisy observation Y^n, the goal is to determine whether X^n is the zero vector or a *sparse* vector, whose support size is approximately $n\epsilon_n$ and non-zero entries are distributed according to P_n. Therefore, the distribution P_n represents the prior knowledge of the signal. The case of P_n being a point mass is treated in [8], [11]. The case of Rademacher P_n in covered in [23, Chapter 8]. The heteroscedastic case where P_n is Gaussian is considered in [6]. These results can be recovered by specializing the general conclusion in the present paper.

Moreover, our results also shed light on what governs the fundamental detection limit in Gaussian noise when the signal does not necessarily have equal strength. For example, consider the classical setup (2) where the signal strength μ_n is now a random variable. If we have $\mu_n = \sqrt{2r\log n}\, X$ for some random variable X, then the resulting detectable region is given by the Ingster-Donoho-Jin expression (20) scaled by the L_∞-norm of X. On the other hand, it is also possible that certain distributions of μ_n induces different shapes of detectable region than Fig. 2. See Sections III-A and V-B for further discussions.

It should be noted that while the setting of the present paper is quite general, the main assumption is that the null distribution is known. This is in contrast with the recent work [24] by Arias-Castro and Wang, which considered unknown null hypotheses and proposed non-parametric tests based on the symmetry of the null and asymmetry of the alternative distributions.

C. Organization

The rest of the paper is organized as follows. Section II states the setup, defines the fundamental limit of sparse mixture detection and reviews some previously known results. The main results of the paper are presented in Sections III and IV, where we provide an explicit characterization of the optimal detection boundary under mild technical conditions. Moreover, it is shown in Section IV that the higher criticism test achieves the optimal performance adaptively. Section V particularizes the general result to various special cases to give explicit formulae of the fundamental limits. Discussions of generalizations and open problems are presented in Section VI. The main theorems are proved in Section VII, while the proofs of the technical lemmas are relegated to the appendices.

D. Notations

Throughout the paper, Φ and φ denote the cumulative distribution function (CDF) and the density of the standard normal distribution respectively. Let $\bar{\Phi} = 1-\Phi$. Let P^n denote the n-fold product measure of P. We say P is absolutely continuous with respect to Q, denoted by $P \ll Q$, if $P(A) = 0$ for any measurable set A such that $Q(A) = 0$. We say P is singular with respect to Q, denoted by $P \perp Q$, if there exists a measurable A such that $P(A) = 1$ and $Q(A) = 0$. We denote $a_n = o(b_n)$ if $\limsup_{n\to\infty} \frac{|a_n|}{|b_n|} = 0$, $a_n = \omega(b_n)$ if $b_n = o(a_n)$, $a_n = O(b_n)$ if $\limsup_{n\to\infty} \frac{|a_n|}{|b_n|} < \infty$ and $a_n = \Omega(b_n)$ if $b_n = O(a_n)$. These asymptotic notations extend naturally to probabilistic setups, denoted by $o_{\mathbb{P}}, \omega_{\mathbb{P}}$, etc., where limits are in the sense of convergence in probability.

II. Fundamental Limits and Characterization

In this section we define the fundamental limits for testing the hypotheses (10) in terms of the sparsity parameter β. An equivalent characterization in terms of the Hellinger distance is also given.

A. Fundamental Limits of Detection

It is easy to see that as the non-null proportion ϵ_n decreases, the signal is more sparse and the testing problem in (10) becomes more difficult. Recall that ϵ_n is given by (3) where $\beta \geq 0$ parametrizes the sparsity level. Thus, the question of detectability boils down to characterizing the smallest (resp.

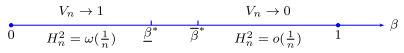

Fig. 1. Critical values of β and regimes of (in)distinguishability of the hypotheses (11) in the large-n limit.

largest) β such that the hypotheses in (10) can be distinguished with probability tending to one (resp. zero), when the sample size n is large.

For testing between two probability measures P and Q, denote the optimal sum of Type-I and Type-II error probabilities by

$$\mathcal{E}(P, Q) \triangleq \inf_A \{P(A) + Q(\circ A)\}, \tag{13}$$

where the infimum is over all measurable sets A. By the Neyman-Pearson Lemma [25], $\mathcal{E}(P, Q)$ is achieved by the likelihood ratio test: declare P if and only if $\frac{\mathrm{d}P}{\mathrm{d}Q} \geq 1$. Moreover, $\mathcal{E}(P, Q)$ can be expressed in terms of the *total variation distance*

$$\mathsf{TV}(P, Q) \triangleq \sup_A |P(A) - Q(A)| = \frac{1}{2} \int |\mathrm{d}P - \mathrm{d}Q| \tag{14}$$

as

$$\mathcal{E}(P, Q) = 1 - \mathsf{TV}(P, Q). \tag{15}$$

For a fixed sequence $\{(Q_n, G_n)\}$, denote the total variation between the null and alternative in (10) by

$$V_n(\beta) \triangleq \mathsf{TV}(Q_n^n,\ ((1 - n^{-\beta})Q_n + n^{-\beta}G_n)^n), \tag{16}$$

which takes values in the unit interval. In view of (15), the fundamental limits of testing the hypothesis (10) are defined as follows.

Definition 1.

$$\underline{\beta}^* \triangleq \sup\{\beta \geq 0 : V_n(\beta) \to 1\}, \tag{17}$$

$$\overline{\beta}^* \triangleq \inf\{\beta \geq 0 : V_n(\beta) \to 0\}. \tag{18}$$

If $\overline{\beta}^* = \underline{\beta}^*$, the common value is denoted by β^*.

As illustrated by Fig. 1, the operational meaning of $\underline{\beta}^*$ and $\overline{\beta}^*$ are as follows: for any $\beta > \overline{\beta}^*$, all sequences of tests have vanishing probability of success; for any $\beta < \underline{\beta}^*$, there exists a sequence of tests with vanishing probability of error. In information-theoretic parlance, if $\overline{\beta}^* = \underline{\beta}^* = \beta^*$, we say *strong converse* holds, in the sense that if $\overline{\beta} > \beta^*$, all tests fail with probability tending to one; if $\beta < \beta^*$, there exists a sequence of tests with vanishing error probability.

Clearly, $\overline{\beta}^*$ and $\underline{\beta}^*$ only depend on the sequence $\{(Q_n, G_n)\}$. The following lemma, proved in Appendix, shows that it is always sufficient to restrict the range of β to the unit interval.

Lemma 1.

$$0 \leq \underline{\beta}^* \leq \overline{\beta}^* \leq 1. \tag{19}$$

In the Gaussian mixture model with $Q_n = \mathcal{N}(0, 1)$, if the sequence $\{G_n\}$ is parametrized by some parameter r, the fundamental limit β^* in Definition 1 is a function of r, denoted by $\beta^*(r)$. For example, in the Ingster-Donoho-Jin setup (2)

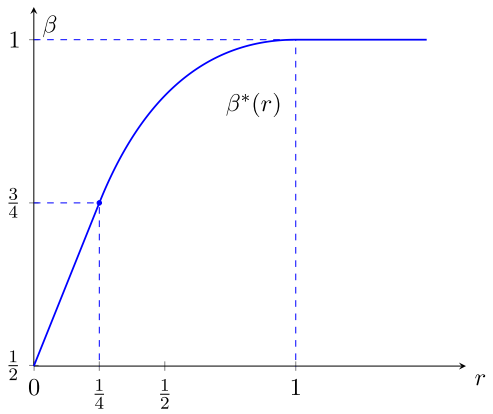

Fig. 2. Ingster-Donoho-Jin detection boundary (20) and the detectable region (below the curve).

where $G_n = \mathcal{N}(\mu_n, 1)$, β^*, denoted by β^*_{IDJ}, can be obtained by inverting (7):

$$\beta^*_{\mathrm{IDJ}}(r) = \begin{cases} \frac{1}{2} + r & 0 < r \leq \frac{1}{4} \\ 1 - (1 - \sqrt{r})_+^2 & r > \frac{1}{4}. \end{cases} \tag{20}$$

In terms of (20), the detectable region is given by the strict hypograph $\{(r, \beta) : \beta < \beta^*_{\mathrm{IDJ}}(r)\}$. The function β^*_{IDJ}, plotted in Fig. 2, plays an important role in our later derivations. Similarly, for the heteroscedastic mixture (8), inverting (9) gives

$$\beta^*(r, \sigma^2) = \begin{cases} \frac{1}{2} + \frac{r}{2-\sigma^2} & 2\sqrt{r} + \sigma^2 \leq 2 \\ 1 - \frac{(1-\sqrt{r})_+^2}{\sigma^2} & 2\sqrt{r} + \sigma^2 > 2. \end{cases} \tag{21}$$

As shown in Section V, all the above results can be obtained in a unified manner as a consequence of the general results in Section III.

B. Equivalent Characterization Via the Hellinger Distance

Closely related to the total variation distance is the Hellinger distance [26, Chapter 2]

$$H^2(P, Q) \triangleq \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}Q})^2,$$

which takes values in the interval $[0, 2]$ and satisfies the following relationship:

$$\frac{1}{2}H^2(P, Q) \leq \mathsf{TV}(P, Q) \leq H(P, Q)\sqrt{1 - \frac{H^2(P, Q)}{4}} \leq 1. \tag{22}$$

Therefore, the total variation distance converges to zero (resp. one) is equivalent to the squared Hellinger distance converges to zero (resp. two). We will be focusing on the Hellinger distance partly due to the fact that it tensorizes under the product measures as follows:

$$H^2(P^n, Q^n) = 2 - 2\left(1 - \frac{H^2(P, Q)}{2}\right)^n. \tag{23}$$

Denote the Hellinger distance between the null and the alternative by

$$H_n^2(\beta) \triangleq H^2(Q_n, (1 - n^{-\beta})Q_n + n^{-\beta}G_n). \tag{24}$$

In view of (17)–(18) and (23), the fundamental limits $\overline{\beta}^*$ and $\underline{\beta}^*$ can be equivalently defined as follows in terms of the asymptotic squared Hellinger distance:

$$\underline{\beta}^* = \sup\left\{\beta \geq 0 : H_n^2(\beta) = \omega(n^{-1})\right\}, \tag{25}$$

$$\overline{\beta}^* = \inf\left\{\beta \geq 0 : H_n^2(\beta) = o(n^{-1})\right\}. \tag{26}$$

III. Main Results

In this section we characterize the detectable region explicitly by analyzing the exact asymptotics of the Hellinger distance induced by the sequence of distributions $\{(Q_n, G_n)\}$.

A. Characterization of β^* for Gaussian Mixtures

This subsection we focus on the case of sparse normal mixture with $Q_n = \mathcal{N}(0,1)$ and G_n absolutely continuous. We will argue in Section III-C that by performing the Lebesgue decomposition on G_n if necessary, we can reduce the general problem to the absolutely continuous case.

We first note that the *essential supremum* of a measurable function f with respect to a measure ν is defined as

$$\operatorname*{ess\,sup}_x f(x) \triangleq \inf\{a \in \mathbb{R} : \nu(\{x : f(x) > a\}) = 0\}.$$

We omit mentioning ν if ν is the Lebesgue measure. Now we are ready to state the main result of this section.

Theorem 1. Let $Q_n = \mathcal{N}(0,1)$. Assume that G_n has a density g_n with respect to the Lebesgue measure. Denote the log-likelihood ratio by

$$\ell_n \triangleq \log \frac{g_n}{\varphi}. \tag{27}$$

Let $\alpha : \mathbb{R} \to \mathbb{R}$ be a measurable function and define

$$\beta^\sharp = \frac{1}{2} + 0 \vee \operatorname*{ess\,sup}_{u\in\mathbb{R}} \left\{\alpha(u) - u^2 + \frac{u^2 \wedge 1}{2}\right\}. \tag{28}$$

1) If

$$\liminf_{n\to\infty} \frac{\ell_n(u\sqrt{2\log n})}{\log n} \geq \alpha(u) \tag{29}$$

uniformly in $u \in \mathbb{R}$, where $\alpha > 0$ on a set of positive Lebesgue measure, then $\underline{\beta}^* \geq \beta^\sharp$.

2) If

$$\limsup_{n\to\infty} \frac{\ell_n(u\sqrt{2\log n})}{\log n} \leq \alpha(u) \tag{30}$$

uniformly in $u \in \mathbb{R}$, then $\overline{\beta}^* \leq \beta^\sharp$.

Consequently, if the limits in (29) and (30) agree and $\alpha > 0$ on a set of positive Lebesgue measure, then $\beta^* = \beta^\sharp$.

Proof: Section VII-B. □

To appreciate the result of Theorem 1, note that the detectability of the sparse mixture in (11) is determined by the non-null effect G_n: The more G_n deviates from the standard normal, the sparser we can afford the alternative to be, and, consequently, the larger the β^* is. The difference between G_n and $\mathcal{N}(0,1)$ relevant to the detection boundary is captured by the function $u \mapsto \alpha(u)$, which is determined by the asymptotic behavior of the log-likelihood ratio ℓ_n evaluated as the n^{-s} and $(1-n^{-s})$-quantile of the null distribution (with $u = \pm\sqrt{s}$). This intuition carries over to the generalization in Section III-B where the null distribution can be non-Gaussian and dependent on the sample size n. Recall from (24) that reliable (resp. unreliable) detection is characterized by the asymptotics of the Hellinger distance $H_n^2(\beta) = \omega(n^{-1})$ (resp. $o(n^{-1})$), whose exponent is dictated by the largest exponent of the integral via a generalized Laplace method. This explains the supremum in the formula (28).

The role of the function α in characterizing the detection limit β^* is fundamental in the following sense. Assuming the setup of Theorem 1, we ask the question in the reverse direction: What kind of function α can arise in equations (29) and (30)? The following lemma (proved in Section VII-B) gives a necessary and sufficient condition for α. Moreover, for every function α satisfying the condition (32) below, there exists a sequence of mixture whose β^* is given by the formula (28). In the special case of convolutional models, the function α needs to satisfy more stringent conditions, which we also discuss below.

Lemma 2. Suppose

$$\lim_{n\to\infty} \frac{\ell_n(u\sqrt{2\log n})}{\log n} = \alpha(u), \tag{31}$$

holds uniformly in $u \in \mathbb{R}$ for some measurable function $\alpha : \mathbb{R} \to \mathbb{R}$. Then for some $t > 0$,

$$\int_{\mathbb{R}} \exp(t(\alpha(u) - u^2))\mathrm{d}u < \infty \text{ and } \operatorname*{ess\,sup}_{u\in\mathbb{R}}\{\alpha(u) - u^2\} = 0 \tag{32}$$

In particular, $\alpha(u) \leq u^2$ Lebesgue-a.e. Conversely, for all measurable α that satisfies (32), there exists a sequence of $\{G_n\}$, such that (31) holds.

Additionally, if the model is convolutional, i.e., $G_n = P_n * \mathcal{N}(0,1)$, then α is convex.

In many applications, we want to know how fast the optimal error probability decays if β lies in the detectable region. The following result gives the precise asymptotics for the Hellinger distance, which also gives upper bounds on the total variation, in view of (22).

Theorem 2. Assume that (31) holds. For any $\beta \geq \frac{1}{2}$, the exponent of the Hellinger distance (24) is given by

$$\lim_{n\to\infty} \frac{\log H_n^2(\beta)}{\log n} = \mathsf{E}(\beta), \tag{33}$$

where

$$\mathsf{E}(\beta) = \operatorname*{ess\,sup}_{u\in\mathbb{R}}\{(2(\alpha(u) - \beta)) \wedge (\alpha(u) - \beta) - u^2\} \tag{34}$$

$$= \operatorname*{ess\,sup}_{u:\alpha(u)\leq\beta} \{2\alpha(u) - 2\beta - u^2\} \vee \operatorname*{ess\,sup}_{u:\alpha(u)>\beta} \{\alpha(u) - \beta - u^2\} \tag{35}$$

which satisfies $\mathsf{E}(\beta) > -1$ (resp. $\mathsf{E}(\beta) < -1$) if and only if $\beta < \beta^\sharp$ (resp. $\beta > \beta^\sharp$).

As an application of Theorem 1, the following result relates the fundamental limit β^* of the convolutional models to the classical Ingster-Donoho-Jin detection boundary:

Corollary 1. Let $G_n = P_n * \mathcal{N}(0,1)$. Assume that P_n has a density p_n which satisfies that

$$\lim_{n\to\infty} \frac{\log p_n(t\sqrt{2\log n})}{\log n} = -f(t) \tag{36}$$

uniformly in $t \in \mathbb{R}$ for some measurable $f : \mathbb{R} \to \mathbb{R}$. Then

$$\beta^* = \operatorname*{ess\,sup}_{t\in\mathbb{R}} \{\beta^*_{\mathrm{IDJ}}(t^2) - f(t)\} \tag{37}$$

where β^*_{IDJ} is the Ingster-Donoho-Jin detection boundary defined in (20).

It should be noted that the convolutional case of the normal mixture detection problem is briefly discussed in [6, Section 6.1], where inner and outer bounds on the detection boundary are given but do not meet. Here Corollary 1 completely settles this question. See Section V for more examples.

We conclude this subsection with a few remarks on Theorem 1.

Remark 1 (Extremal cases). Under the assumption that the function $\alpha > 0$ on a set of positive Lebesgue measure, the formula (28) shows that the fundamental limit β^* lies in the very sparse regime ($\frac{1}{2} \le \beta^* \le 1$). We discuss the two extremal cases as follows:

1) *Weak signal*: Note that $\beta^* = \frac{1}{2}$ if and only if $\alpha(u) \le u^2 - \frac{u^2\wedge 1}{2}$ almost everywhere. In this case the non-null effect is too weak to be detected for any $\beta > \frac{1}{2}$. One example is the zero-mean heteroscedastic case $G_n = \mathcal{N}(0,\sigma^2)$ with $\sigma^2 \le 2$. Then we have $\alpha(u) \le \frac{u^2}{2}$.
2) *Strong signal*: Note that $\beta^* = 1$ if and only if there exists u, such that $|u| \ge 1$ and

$$\alpha(u) = u^2. \tag{38}$$

At this particular u, the density of the signal satisfies $g_n(u\sqrt{2\log n}) = n^{-o(1)}$, which implies that there exists significant mass beyond $\sqrt{2\log n}$, the extremal value under the null hypothesis [27]. This suggests the possibility of constructing test procedures based on the *sample maximum*. Indeed, to understand the implication of (38) more quantitatively, let us look at an even weaker condition: there exists u such that $|u| \ge 1$ and

$$\limsup_{n\to\infty} \frac{1}{\log n} \log \frac{1}{\mathbb{P}\left\{u^{-1}Y_n \ge \sqrt{2\log n}\right\}} = 0, \tag{39}$$

which, as shown in Appendix , implies that $\beta^* = 1$.

Remark 2. In general β^* need not exist. Based on Theorem 1, it is easy to construct a Gaussian mixture where $\overline{\beta}^*$ and $\underline{\beta}^*$ do not coincide. For example, let α_0 and α_1 be two measurable functions which satisfy Lemma 2 and give rise to different values of $\beta^\sharp$ in (28), which we denote by $\beta_0^\sharp < \beta_1^\sharp$. Then there exist sequences of distributions $\{G_n^{(0)}\}$ and $\{G_n^{(1)}\}$ which satisfy (31) for α_0 and α_1 respectively. Now define $\{G_n\}$ by $G_{2k} = G_k^{(0)}$ and $G_{2k+1} = G_k^{(1)}$. Then by Theorem 1, we have $\underline{\beta}^* = \beta_0^\sharp < \overline{\beta}^* = \beta_1^\sharp$.

B. Non-Gaussian Mixtures

The detection boundary in [8], [11] is obtained by deriving the limiting distribution of the log-likelihood ratio which relies on the normality of the null hypothesis. In contrast, our approach is based on analyzing the sharp asymptotics of the Hellinger distance. This method enables us to generalize the result of Theorem 1 to sparse non-Gaussian mixtures (10), where we even allow the null distribution Q_n to vary with the sample size n.

Theorem 3. Consider the hypothesis testing problem (10). Let $G_n \ll Q_n$. Denote by F_n and z_n the CDF and the quantile function of G_n, respectively, i.e.,

$$z_n(p) = \inf\{y \in \mathbb{R} : F_n(y) \ge p\}. \tag{40}$$

Assume that the log-likelihood ratio

$$\ell_n = \log \frac{\mathrm{d}G_n}{\mathrm{d}Q_n} \tag{41}$$

satisfies

$$\lim_{n\to\infty} \sup_{s\ge(\log_2 n)^{-1}} \left| \frac{\ell_n(z_n(n^{-s})) \vee \ell_n(z_n(1-n^{-s}))}{\log n} - \gamma(s) \right| = 0 \tag{42}$$

as $n \to \infty$ uniformly in $s \in \mathbb{R}_+$ for some measurable function $\gamma : \mathbb{R}_+ \to \mathbb{R}$. If $\gamma > 0$ on a set of positive Lebesgue measure, then

$$\beta^* = \frac{1}{2} + 0 \vee \operatorname*{ess\,sup}_{s\ge 0} \left\{ \gamma(s) - s + \frac{s\wedge 1}{2} \right\}. \tag{43}$$

The function γ appearing in Theorem 3 plays the same role as the function α in Theorem 1 for the Gaussian case, which is also determined by the log-likelihood ratio evaluated at the n^{-s} and $(1-n^{-s})$-quantiles of the null distribution for $s > 0$. Indeed, comparing Theorem 3 with Theorem 1, we see that the uniform convergence condition (31) is naturally replaced by the uniform convergence of the log-likelihood ratio evaluated at the null quantile. Using the fact that $\frac{z}{1+z^2} \le \frac{\bar\Phi(z)}{\varphi(z)} \le \frac{1}{z}$ for all $z > 0$ [28, 7.1.13], which implies that

$$\bar\Phi(z) = \frac{\varphi(z)}{z}(1+o(1)) \tag{44}$$

as $z \to \infty$, we can recover Theorem 1 from Theorem 3 by setting $\gamma(s) = \alpha(-\sqrt{s}) \vee \alpha(\sqrt{s})$. Analogously, the function γ satisfies the same condition as in Lemma 2.

C. Decomposition of the Alternative

The results in Theorem 1 and Theorem 3 are obtained under the assumption that the non-null effect G_n is absolutely continuous with respect to the null distribution Q_n. Next we show that it does not lose generality to focus our attention on this case. Using the Hahn-Lebesgue decomposition [29, Theorem 1.6.3], we can write

$$G_n = (1-\kappa_n)G_n' + \kappa_n \nu_n \tag{45}$$

for some $\kappa_n \in [0,1]$, where $G_n' \ll Q_n$ and $\nu_n \perp Q_n$. Put

$$\epsilon_n' = \frac{\epsilon_n(1-\kappa_n)}{1-\epsilon_n\kappa_n} \quad \text{and} \quad Q_n' = (1-\epsilon_n')Q_n + \epsilon_n' G_n', \tag{46}$$

which satisfies $Q_n' \ll Q_n$. Then $(1-\epsilon_n)Q_n + \epsilon_n G_n = (1-\epsilon_n\kappa_n)Q_n' + \epsilon_n\kappa_n\nu_n$. By Lemma 7,

$$H^2(Q_n, (1-\epsilon_n)Q_n + \epsilon_n G_n) = \Theta(\epsilon_n\kappa_n \vee H^2((1-\epsilon')Q_n + \epsilon_n' G_n')) \quad (47)$$

Therefore the asymptotic Hellinger distance of the original problem is completely determined by $\epsilon_n\kappa_n$ and the square-Hellinger distance $H^2((1-\epsilon_n')Q_n + \epsilon_n' G_n')$, which is also of a *sparse mixture* form, with (ϵ_n, G_n) replaced by (ϵ_n', G_n') given in (46). In particular, we note the following special cases.

1) If $\epsilon_n\kappa_n = O(n^{-1})$, then $H^2(Q_n, (1-\epsilon_n)Q_n + \epsilon_n G_n) = o(n^{-1})$ (resp. $\omega(n^{-1})$) if and only if $H^2(Q_n, (1-\epsilon_n')Q_n + \epsilon_n' G_n') = o(n^{-1})$ (resp. $\omega(n^{-1})$), which means that detectability of the original sparse mixture coincide with the new mixture.
2) If $\epsilon_n\kappa_n = \omega(n^{-1})$, then $H^2(Q_n, (1-\epsilon_n)Q_n + \epsilon_n G_n) = \omega(n^{-1})$, which means that the original sparse mixture can be detected reliably. In fact, a trivial optimal test is to reject the null hypothesis if there exists one sample lying in the support of the singular component ν_n.

IV. Adaptive Optimality of Higher Criticism Tests

As discussed in Section II-A, the fundamental limit β^* of testing sparse normal mixtures (11) can be achieved by the likelihood ratio test. However, in general the likelihood ratio test requires the knowledge of the alternative distribution, which is typically not accessible in practice. To overcome this limitation, it is desirable to construct *adaptive* testing procedures to achieve the optimal performance simultaneously for a collection of alternatives. The basic idea of adaptive procedures usually involves comparing the empirical distribution of the data to the null distribution, which is assumed to be known.

For the problem of detecting sparse normal mixtures, it is especially relevant to construct adaptive procedures, since in practice the underlying sparsity level and the non-zero priors are usually unknown. Toward this end, Donoho and Jin [8] introduced an adaptive test based on Tukey's *higher criticism* statistic. For the special case of (2), i.e., $P_n = \delta_{\sqrt{2r\log n}}$, it is shown that the higher criticism test achieves the optimal detection boundary (20) while being adaptive to the unknown non-null parameters (β, r). Following the generalization by Jager and Wellner [17] via Rényi divergence, next we explain briefly the gist of the higher criticism test.

Given the data $Y_1, \ldots, Y_n$, denote the empirical CDF by

$$\mathbb{F}_n(t) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{Y_i \le t\}},$$

respectively. Similar to the Kolmogorov-Smirnov statistic [30, p. 91] which computes the L_∞-distance (maximal absolute difference) between the empirical CDF and the null CDF, the higher criticism statistic is the maximal pointwise χ^2-divergence between the null and the empirical CDF. We first introduce a few auxiliary notations. Recall that the χ^2-divergence between two probability measures is defined as

$$\chi^2(P \,||\, Q) \triangleq \int \left(\frac{\mathrm{d}P}{\mathrm{d}Q} - 1\right)^2 \mathrm{d}Q.$$

In particular, the binary χ^2-divergence function (i.e., the χ^2-divergence between Bernoulli distributions) is given by

$$\chi^2(\mathrm{Bern}(p) \,||\, \mathrm{Bern}(q)) = \frac{(p-q)^2}{q(1-q)},$$

where $\mathrm{Bern}(p)$ denotes the Bernoulli distribution with bias p. The higher criticism statistic is defined by

$$\mathrm{HC}_n \triangleq \sup_{t\in\mathbb{R}} \sqrt{n\chi^2(\mathrm{Bern}(\mathbb{F}_n(t)) \,||\, \mathrm{Bern}(\Phi(t)))} \quad (48)$$

$$= \sqrt{n}\sup_{t\in\mathbb{R}} \frac{|\mathbb{F}_n(t) - \Phi(t)|}{\sqrt{\Phi(t)\bar{\Phi}(t)}} \quad (49)$$

Based on the statistics (48), the higher criticism test declares H_1 if and only if

$$\mathrm{HC}_n > \sqrt{2(1+\delta)\log\log n} \quad (50)$$

where $\delta > 0$ is an arbitrary fixed constant.

The next result shows that the higher criticism test achieves the fundamental limit β^* characterized by Theorem 1 while being adaptive to all sequences of distributions $\{G_n\}$ which satisfy the regularity condition (31). This result generalizes the adaptivity of the higher criticism procedure far beyond the original equal-signal-strength setup in [8] and the heteroscedastic extension in [6].

Theorem 4. Under the same assumption of Theorem 1, for any $\beta > \beta^*$, the sum of Type-I and Type-II error of the higher criticism test (50) vanishes as $n \to \infty$.

V. Examples

In this section we particularize the general result in Theorem 1 to several interesting special cases to obtain explicit detection boundaries.

A. Ingster-Donoho-Jin Detection Boundary

We derive the classical detection boundary (20) from Theorem 1 for the equal-signal-strength setup (2), which is a convolutional model with signal distribution

$$P_n = \delta_{\mu_n} \quad (51)$$

and μ_n in (4). The log-likelihood ratio is given by

$$\ell_n(y) = \log\frac{\varphi(y-\mu_n)}{\varphi(y)} = -\frac{\mu_n^2}{2} + \mu_n y = -r\log n + \sqrt{2r\log n}\, y.$$

Plugging in $y = u\sqrt{2\log n}$, we have $\ell_n(u\sqrt{2\log n}) = -r\log n + 2u\sqrt{r}\log n$. Consequently, the condition (31) is fulfilled uniformly in $u \in \mathbb{R}$ with

$$\alpha(u) = 2u\sqrt{r} - r. \quad (52)$$

Straightforward calculation yields that

$$\operatorname*{ess\,sup}_{u\in\mathbb{R}} \left\{2u\sqrt{r} - r - u^2 + \frac{u^2\wedge 1}{2}\right\} = \begin{cases} r & 0 < r \le \frac{1}{4} \\ \frac{1}{2} - (1-\sqrt{r})_+^2 & r > \frac{1}{4}. \end{cases} \quad (53)$$

Applying Theorem 1, we obtain the desired expression (20) for $\beta^*(r)$.

As a variation of (51), the symmetrized version of (51)

$$P_n = \frac{1}{2}(\delta_{\mu_n} + \delta_{-\mu_n}) \tag{54}$$

was considered in [23, Section 8.1.6], whose detection boundary is shown to be identical to (20). Indeed, for binary-valued signal distributed according to (54), we have

$$\begin{aligned}\ell_n(u\sqrt{2\log n}) &= -\frac{\mu_n^2}{2} + \log\cosh(\mu_n u\sqrt{2\log n}) \\ &= -r\log n + \log(n^{2u\sqrt{r}} + n^{-2u\sqrt{r}}) - \log 2\end{aligned}$$

which gives rise to

$$\alpha(u) = 2|u|\sqrt{r} - r \tag{55}$$

Comparing (55) with (52) and (53), we conclude that the detection boundary (20) still applies.

B. Dilated Signal Distributions

Generalizing both the unary and binary signal distributions in Section V-A, we consider P_n that is the distribution of the random variable

$$X_n = \mu_n X \tag{56}$$

where $\mu_n > 0$ is a sequence of positive numbers and X is distributed according to a fixed distribution P, parameterizing the shape of the signal. In other words, P_n is the dilation of P by μ_n. We ask the following question: By choosing the sequence μ_n and the random variable X, is it possible to have detection boundaries which are shaped differently than the classical Ingster-Donoho-Jin detection boundary?

It turns out that for $\mu_n = \sqrt{2\log n}$, the answer to the above question is negative. As the next theorem shows, the detection boundary is given by that of the classical setup rescaled by the L_∞-norm of X. Note that (51) and (54) corresponds to $P = \delta_{\sqrt{r}}$ and $P = \frac{1}{2}(\delta_{\sqrt{r}} + \delta_{-\sqrt{r}})$, respectively.

Corollary 2. Consider the convolutional model $G_n = P_n * \mathcal{N}(0,1)$, where P_n is the distribution of $\sqrt{2\log n}X$. Then

$$\beta^* = \beta^*_{\mathrm{IDJ}}(\|X\|_\infty^2) = \begin{cases} \|X\|_\infty^2 + \frac{1}{2} & 0 < \|X\|_\infty \le \frac{1}{2} \\ 1 - (1 - \|X\|_\infty)_+^2 & \|X\|_\infty > \frac{1}{2}. \end{cases} \tag{57}$$

Proof: Recall that $\beta^*_{\mathrm{IDJ}}(\cdot)$ denotes the Ingster-Donoho-Jin detection boundary defined in (20). Since the log-likelihood ratio is given by $\ell_n(y) = \mathbb{E}\left[\exp(-\frac{X_n^2}{2} + X_n y)\right]$, we have

$$\begin{aligned}\ell_n(u\sqrt{2\log n}) &= \log \mathbb{E}\left[n^{-X^2+2uX}\right] \\ &= \operatorname*{ess\,sup}_X \left\{-X^2 + 2uX\right\}\log n(1+o(1)), \end{aligned} \tag{58}$$

where we have applied Lemma 3 and the essential supremum in (58) is with respect to P, the distribution of X. Therefore $\alpha(u) = \operatorname{ess\,sup}_X \left\{-X^2 + 2uX\right\}$. Applying Theorem 1 yields the existence of β^*, given by

$$\begin{aligned}\beta^* &= \frac{1}{2} + \operatorname*{ess\,sup}_{u\in\mathbb{R}} \left\{ \operatorname*{ess\,sup}_X \left\{ -X^2 + 2uX \right\} - u^2 + \frac{u^2 \wedge 1}{2} \right\} \\ &= \frac{1}{2} + \operatorname*{ess\,sup}_X \operatorname*{ess\,sup}_{u\in\mathbb{R}} \left\{ -X^2 + 2uX - u^2 + \frac{u^2 \wedge 1}{2} \right\} \\ &= \operatorname*{ess\,sup}_X \beta^*_{\mathrm{IDJ}}(X^2) \\ &= \beta^*_{\mathrm{IDJ}}(\|X\|_\infty^2), \end{aligned} \tag{59}$$

where (59) follows from the facts that $\beta^*_{\mathrm{IDJ}}(\cdot)$ is increasing and that $\|X\|_\infty = \operatorname{ess\,sup}|X|$. □

Remark 3. Corollary V-B tightens the bounds given at the end of [6, Section 6.1] based on the interval containing the signal support. From (57) we see that the detection boundary coincides with the classical case with $\sqrt{r}$ replaced by L_∞-norm of X. Therefore, as far as the detection boundary is concerned, only the support of X matters and the detection problem is driven by the maximal signal strength. In particular, for $\|X\|_\infty \ge 1$ or non-compactly supported X, we obtain the degenerate case $\beta^* = 1$ (see also Remark 1 about the strong-signal regime). However, it is possible that the density of X plays a role in finer asymptotics of the testing problem, e.g., the convergence rate of the error probability and the limiting distribution of the log-likelihood ratio at the detection boundary.

One of the consequences of Corollary 2 is the following: as long as $\mu_n = \sqrt{2\log n}$, non-compactly supported X results in the degenerate case of $\beta^* = 1$, since the signal is too strong to go undetected. However, this conclusion need not be true if μ_n behaves differently. We conclude this subsection by constructing a family of distributions of X with unbounded support and an appropriately chosen sequence $\{\mu_n\}$, such that the detection boundary is non-degenerate: Let X be distributed according to the following *generalized Gaussian* (Subbotin) distribution P_τ [31] with shape parameter $\tau > 0$, whose density is

$$p_\tau(x) = \frac{\tau}{2\Gamma(\tau)}\exp(-|x|^\tau). \tag{60}$$

Put $\mu_n = \sqrt{2r}(\log n)^{\frac{1}{2}-\frac{1}{\tau}}$. Then the density of X_n is given by $v_n(x) = \frac{1}{\mu_n}p(\frac{x}{\mu_n})$. Hence

$$v_n(t\sqrt{2\log n}) = \frac{\tau}{2\Gamma(\tau)\mu_n} n^{-|t|^\tau r^{-\frac{\tau}{2}}},$$

which satisfies the condition (36) with $f(t) = |t|^\tau r^{-\frac{\tau}{2}}$. Applying Corollary 1, we obtain the detection boundary β^* (a two-dimensional *surface* parametrized by (r,τ) shown in Fig. 3) as follows

$$\beta^* = \sup_{t\in\mathbb{R}}\{\beta^*_{\mathrm{IDJ}}(t^2) - |t|^\tau r^{-\frac{\tau}{2}}\} = \sup_{z\ge 0}\{\beta^*_{\mathrm{IDJ}}(rz^2) - z^\tau\} \tag{61}$$

where (20) is the Ingster-Donoho-Jin detection boundary.

Equation (61) can be further simplified for the following special cases.

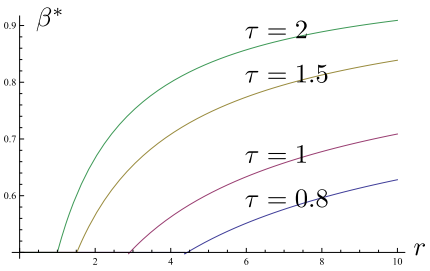

Fig. 3. Detection boundary β^* given by (61) as a function of r for various values of τ.

- $\tau = 1$ (Laplace): Plugging (20) into (61), straightforward computation yields

$$\beta^* = \frac{1}{2} \vee \left(1 - \frac{1}{2\sqrt{r}}\right)_+^2 = \begin{cases} \left(1 - \frac{1}{2\sqrt{r}}\right)^2 & r > \frac{3}{2} + \sqrt{2} \\ \frac{1}{2} & r \le \frac{3}{2} + \sqrt{2}. \end{cases}$$

- $\tau = 2$ (Gaussian): In this case we have $X \sim \mathcal{N}(0, \frac{1}{2})$ and $X_n \sim \mathcal{N}(0, r)$. This is a special case of the heteroscedastic case in [6], which will be discussed in detail in Section V-C. Simplifying (61) we obtain

$$\beta^* = \frac{1}{2} \vee \frac{r}{1+r},$$

which coincides with (67).

C. Heteroscedastic Normal Mixture

The heteroscedastic normal mixtures considered in (8) corresponds to

$$G_n = \mathcal{N}(\mu_n, \sigma^2)$$

with μ_n given in (4) and $\sigma^2 \ge 0$. In particular, if $\sigma^2 \ge 1$, G_n is given by the convolution $G_n = \Phi * P_n$, where the Gaussian component $P_n = \mathcal{N}(\mu_n, \sigma^2 - 1)$ models the variation in the signal amplitude.

For any $u \in \mathbb{R}$,

$$\ell_n(u\sqrt{2\log n}) = \log \frac{\varphi\left(\frac{u\sqrt{2\log n} - \mu_n}{\sigma}\right)}{\varphi(u\sqrt{2\log n})} = \alpha(u)\log n, \quad (62)$$

where

$$\alpha(u) = u^2 - \frac{(u - \sqrt{r})^2}{\sigma^2}.$$

Similar to the calculation in Section V-A, we have[1]

$$\sup_{0 \le s \le 1} \left\{\alpha(s) - \frac{s}{2}\right\} = \begin{cases} \frac{r}{2-\sigma^2} & 2\sqrt{r} + \sigma^2 \le 2 \\ \frac{1}{2} - \frac{(1-\sqrt{r})_+^2}{\sigma^2} & 2\sqrt{r} + \sigma^2 > 2 \end{cases} \quad (63)$$

and

$$\sup_{s \ge 1} \{\alpha(s) - s\} = -\frac{(1 - \sqrt{r})_+^2}{\sigma^2}. \quad (64)$$

Note that $\frac{r}{2-\sigma^2} - (\frac{1}{2} - \frac{(1-\sqrt{r})_+^2}{\sigma^2}) \ge \frac{(\sigma^2 + 2\sqrt{2} - 2)^2}{2\sigma^2(2-\sigma^2)} \ge 0$ if $2\sqrt{r} + \sigma^2 \le 2$. Assembling (63)–(64) and applying Theorem 1, we have

$$\beta^*(r, \sigma^2) = \frac{1}{2} + \left(\frac{r}{2-\sigma^2}\right) \vee \left(\frac{1}{2} - \frac{(1-\sqrt{r})_+^2}{\sigma^2}\right) \quad (65)$$

$$= \begin{cases} \frac{1}{2} + \frac{r}{2-\sigma^2} & 2\sqrt{r} + \sigma^2 \le 2 \\ 1 - \frac{(1-\sqrt{r})_+^2}{\sigma^2} & 2\sqrt{r} + \sigma^2 > 2. \end{cases} \quad (66)$$

Solving the equation $\beta^*(r, \sigma^2) = \beta$ in r yields the equivalent detection boundary (9) in terms of r. In the special case of $r = 0$, where the signal is distributed according to $P_n = \mathcal{N}(0, \tau^2)$, we have

$$\beta^*(0, 1 + \tau^2) = \frac{\tau^2 \vee 1}{1 + \tau^2 \vee 1}. \quad (67)$$

Therefore, as long as the signal variance exceeds that of the noise, reliable detection is possible in the very sparse regime $\beta > \frac{1}{2}$, even if the average signal strength does not tend to infinity.

D. Non-Gaussian Mixtures

We consider the detection boundary of the following generalized Gaussian location mixture which was studied in [8, Section 5.2]:

$$H_0^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} P_\tau(\cdot) \quad (68)$$

$$\text{v.s.} \quad H_1^{(n)} : Y_i \overset{\text{i.i.d.}}{\sim} (1 - \epsilon_n) P_\tau(\cdot) + \epsilon_n P_\tau(\cdot - \mu_n)$$

where P_τ is defined in (60), and $\mu_n = (r \log n)^{\frac{1}{\tau}}$. Since $z(1 - n^{-s}) = z(n^{-s}) = (s \log n)^{\frac{1}{\tau}}(1 + o(1))$ uniformly in s, (42) is fulfilled with $\gamma(s) = s - |s^{\frac{1}{\tau}} - r^{\frac{1}{\tau}}|$. Applying Theorem 3, we have

$$\beta^*(r) = \frac{1}{2} + 0 \vee \sup_{s \ge 0} \left(-|s^{\frac{1}{\tau}} - r^{\frac{1}{\tau}}| + \frac{s \wedge 1}{2}\right)$$

$$= \frac{1}{2} + 0 \vee \sup_{u \ge 0} \left(-|u - r^{\frac{1}{\tau}}| + \frac{u^\tau \wedge 1}{2}\right)$$

$$= \begin{cases} 1 & r > 1 \\ \frac{1+r}{2} & \tau \le 1, r \le 1, \\ \frac{1}{2} + \frac{\frac{1}{2} - 2^{\frac{\tau}{1-\tau}}}{(1 - 2^{\frac{1}{1-\tau}})^\tau} r & \tau \ge 1, r < (1 - 2^{\frac{1}{1-\tau}})^\tau \\ 1 - (1 - r^{\frac{1}{\tau}})^\tau & \tau \ge 1, r \ge (1 - 2^{\frac{1}{1-\tau}})^\tau. \end{cases} \quad (69)$$

It is easy to verify that (69) agrees with the results in [8, Theorem 5.1]. Similarly, the detection boundary for exponential-χ_2^2 mixture in [8, Theorem 1.7] can also be derived from Theorem 3.

VI. Discussion

By analyzing the sharp asymptotics of the Hellinger distance, we obtained explicit expressions for the detection boundary for sparse normal mixture in Theorem 1 and generalized the results to non-Gaussian mixtures in Theorem 3 under more technical conditions. This method is in contrast with the approach in [8], [23], which determine the detection boundary by deriving the limiting distribution of the log-likelihood ratio and relies on the normality of the null hypothesis, or the

[1] In the first case of (63) it is understood that $\frac{0}{0} = 0$.

method of analyzing the truncated χ^2-divergence (see, e.g., [11], [24], [32]). In view of the discussion in Section II-B, the limiting behavior of Hellinger distance is an equivalent characterization for the asymptotic (in)distinguishability of the hypotheses. It is interesting to find other applications in high-dimensional detection problems where our Hellinger-based approach is applicable.

While we have focused on a Bayesian setting with independent samples, interesting future directions to pursue include generalizations to a) minimax detection problems such as [20]–[22] where the data are independent but the parameter set are structured, and b) detection with correlated data such as [19] where the lower bound is also obtained by analyzing the Hellinger distance.

We conclude the paper with a few discussions and open problems.

A. Moderately Sparse Regime $0 \le \beta \le \frac{1}{2}$

Our main results in Section III only concern the *very sparse* regime $\frac{1}{2} < \beta < 1$. This is because under the assumption in Theorem 1 that $\alpha > 0$ on a set of positive Lebesgue measure, we always have $\beta^* \ge \frac{1}{2}$. One of the major distinctions between the very sparse and moderately sparse regimes is the effect of symmetrization. To illustrate this point, consider the sparse normal mixture model (11). Given any G_n, replacing it by its symmetrized version $\tilde{G}_n(\mathrm{d}x) \triangleq \frac{G_n(\mathrm{d}x)+G_n(-\mathrm{d}x)}{2}$ always increases the difficulty of testing. This follows from the inequality $H^2(\tilde{G}_n, \Phi) \le H^2(G_n, \Phi)$, a consequence of the convexity of the squared Hellinger distance and the symmetry of Φ. A natural question is: Does symmetrization always have an impact on the detection boundary? In the very sparse regime, it turns out that under the regularity conditions imposed in Theorem 1, symmetrization does not affect the fundamental limit β^*, because both G_n and $\tilde{G}_n$ give rise to the same function α. It is unclear whether $\overline{\beta}^*$ and $\underline{\beta}^*$ remain unchanged if an arbitrary sequence $\{G_n\}$ is symmetrized. However, in the moderately sparse regime, an asymmetric non-null effect can be much more detectable than its symmetrized version. For instance, direct calculation (see for example [6, Section 2.2]) shows that $\beta^*(r) = \frac{1}{2} - r$ for $G_n = \delta_{n^{-r}}$, but $\beta^*(r) = \frac{1}{2} - 2r$ for $G_n = \frac{1}{2}(\delta_{n^{-r}} + \delta_{-n^{-r}})$.

Moreover, unlike in the very sparse regime, moment-based tests can be powerful in the moderately sparse regime, which guarantee that $\overline{\beta}^* \ge \frac{1}{2}$. For instance, in the above examples $G_n = \delta_{n^{-r}}$ or $G_n = \frac{1}{2}(\delta_{n^{-r}} + \delta_{-n^{-r}})$, the detection boundary can be obtained by thresholding the sample mean or sample variance respectively. More sophisticated moment-based tests such as the excess kurtosis tests have been studied in the context of sparse mixtures [4]. It is unclear whether they are always optimal when $\beta < \frac{1}{2}$.

B. Adaptive Optimality of Higher Criticism Tests

While Theorem 4 establishes the adaptive optimality of the higher criticism test in the very sparse regime $\beta > \frac{1}{2}$, the optimality of the higher criticism test in the moderately sparse case $\beta < \frac{1}{2}$ remains an open question. Note that in the classical setup (2), it has been shown [6] that the higher criticism test achieves adaptive optimality for $\beta \in [0, \frac{1}{2}]$ and $\mu_n = n^{-r}$. In this case since $\mu_n = o(1)$, we have $\alpha \equiv 0$ and Theorem 1 thus does not apply. It is possible to obtain a counterpart of Theorem 1 and an analogous expression for β^* for the moderately sparse regime if one assumes a similar uniform approximation property of the log-likelihood ratio, for example, $\ell_n(u\sqrt{\log n}) = n^{-\alpha(u)+o(1)}$ for some function α. Another interesting problem is to investigate the optimality of procedures introduced in [17] based on Rényi divergence under the same setup of Theorem 4.

The optimality of the higher criticism test is established for Gaussian null hypothesis. It is of interests to investigate whether higher criticism test achieves the optimal detection boundary in the non-Gaussian case under the assumption of Theorem 3, or the setting in [24] where the null distribution is unknown.

C. Connections to the Settings of Asymptotic Efficiency

The detection problem considered in this paper is, in a broad sense, related to the concept of asymptotic efficiency in the classical statistics literature. For a test ψ which tests a null hypothesis $H_0 : \theta = \theta_0$ at the significance level a, denote its power at a given alternative θ by $b = P_\theta(\psi = 1)$. Relative efficiency measures such as the Hodges-Lehmann efficiency, Pitman efficiency, and Bahadur efficiency quantify the relative performance of two testing procedures for a fixed level a as $b \to 1$ (Hodges-Lehmann), a fixed power b as $a \to 0$ (Bahadur), or fixed a and b (Pitman), by comparing the required sample sizes for the two tests. See, for example, [33]–[35]. For the detection problem considered in the present paper, we characterize conditions on the parameter values under which there is a detector (test) such that both $a \to 0$ and $b \to 1$. In some broad sense, this complements the three settings considered in the classical asymptotic efficiency literature, which hold at least one of the two values a and b fixed.

VII. Proofs

A. Auxiliary Results

Laplace's method (see, e.g., [36, Section 2.4]) is a technique for analyzing the asymptotics of integrals of the form $\int \exp(Mf)\mathrm{d}\nu$ when M is large. The proof of Theorem 1 uses the following first-order version of the Laplace's method. Since we are only interested in the exponent (i.e., the leading term), we do not use saddle-point approximation in the usual Laplace's method and impose *no* regularity conditions on the function f except for the finiteness of the integral. Moreover, the exponent only depends on the *essential supremum* of f with respect to ν, which is invariant if f is modified on a ν-negligible set.

Lemma 3. Let $(X, \mathcal{F}, \nu)$ be a measure space. Let $F : X \times \mathbb{R}_+ \to \mathbb{R}_+$ be measurable. Assume that

$$\lim_{M\to\infty} \frac{\log F(x, M)}{M} = f(x) \tag{70}$$

holds uniformly in $x \in X$ for some measurable $f : X \to \mathbb{R}$. If $\int_X \exp(M_0 f)\mathrm{d}\nu < \infty$ for some $M_0 > 0$, then

$$\lim_{M\to\infty} \frac{1}{M} \log \int_X F(x, M)\mathrm{d}\nu = \operatorname*{ess\,sup}_{x\in X} f(x). \tag{71}$$

Proof: First we deal with the case of $\operatorname{ess\,sup} f = \infty$, which implies that $\nu(\{f > a\}) > 0$ for all $a > 0$. Moreover, by Chernoff bound, $\nu(\{f > a\}) < \exp(-M_0 a) \int \exp(M_0 f) \mathrm{d}\nu < \infty$. By (70), for any $\epsilon > 0$, there exists $K > M_0$ such that

$$\exp(M(f(x) - \epsilon)) \le F(x, M) \le \exp(M(f(x) + \epsilon)) \quad (72)$$

for all $x \in X$ and $M \ge K$. Therefore,

$$\begin{aligned} \int F(x, M) \mathrm{d}\nu &\ge \exp(-M\epsilon) \int \exp(Mf) \mathrm{d}\nu \\ &\ge \exp(M(a - \epsilon)) \nu(\{f > a\}) \end{aligned}$$

for any $M > 0$ and $a > 0$. Then

$$\liminf_{M \to \infty} \frac{1}{M} \log \int \exp(Mf) \mathrm{d}\nu \ge a - \epsilon.$$

Hence $\lim_{M \to \infty} \frac{1}{M} \log \int \exp(Mf) \mathrm{d}\nu = \infty$, by the arbitrariness of a and ϵ.

Next we assume that $\operatorname{ess\,sup} f < \infty$. By replacing f with $f - \operatorname{ess\,sup} f$, we can assume that $\operatorname{ess\,sup} f = 0$ without loss of any generality. Then $f \le 0$ ν-a.e. Hence, by (72),

$$\begin{aligned} \int F(x, M) \mathrm{d}\nu &\le \int \exp(M(f + \epsilon)) \mathrm{d}\nu \\ &\le \exp(M\epsilon) \int \exp(M_0 f) \mathrm{d}\nu < \infty \end{aligned}$$

holds for all $M \ge K$. By the arbitrariness of ϵ, we have

$$\limsup_{M \to \infty} \frac{1}{M} \log \int \exp(Mf) \mathrm{d}\nu \le 0.$$

For the lower bound, note that, by the definition of $\operatorname{ess\,sup} f = 0$, $\nu(\{f > -\delta\}) > 0$ for all $\delta > 0$. Therefore, by (72), we have

$$\begin{aligned} \int F(x, M) \mathrm{d}\nu &\ge \exp(-M\epsilon) \int \exp(Mf) \mathrm{d}\nu \\ &\ge \exp(-M(\delta + \epsilon)) \nu(\{f > -\delta\}) \end{aligned}$$

for any $M > 0$ and $\delta > 0$. First sending $M \to \infty$ then $\delta \downarrow 0$ and $\epsilon \downarrow 0$, we have

$$\liminf_{M \to \infty} \frac{1}{M} \log \int \exp(Mf) \mathrm{d}\nu \ge 0,$$

completing the proof of (71). □

The following lemma is useful for analyzing the asymptotics of Hellinger distance.

Lemma 4. 1) For any $b > 0$, the function $s \mapsto (\sqrt{1 + b(s-1)} - 1)^2$ is strictly convex on $\mathbb{R}_+$ and strictly decreasing and increasing on $[0, 1]$ and $[1, \infty)$, respectively.

2) For any $t \ge 0$,

$$(\sqrt{2} - 1)^2 t \wedge t^2 \le (\sqrt{1 + t} - 1)^2 \le t \wedge t^2. \quad (73)$$

Proof:

1) Since $t \mapsto \sqrt{1 + t}$ is strictly concave, $s \mapsto (\sqrt{1 + b(s-1)} - 1)^2 = 2 + b(s-1) - 2\sqrt{b(s-1)}$ is strictly convex. Solving for the stationary point yields the minimum at $s = 1$.

2) First we consider $t \ge 1$. Since $t \mapsto (\sqrt{1 + t} - 1)^2 = t - 2\sqrt{1 + t}$ is convex, $t \mapsto \frac{(\sqrt{1+t}-1)^2}{t}$ is increasing. Consequently, we have $(\sqrt{2} - 1)^2 \le \frac{(\sqrt{1+t}-1)^2}{t} \le 1$ for all $t \in [1, \infty)$.

Next we consider $0 \le t \le 1$. By the concavity of $t \mapsto \sqrt{1 + t}$, $t \mapsto \frac{\sqrt{1+t}-1}{t}$ is decreasing. Hence $\sqrt{2} - 1 \le \frac{\sqrt{1+t}-1}{t} \le \frac{1}{2}$ for all $t \in [0, 1]$. Assembling the above two cases yields (73). □

The following lemmas are useful in proving Theorem 4.

Lemma 5. Let $f : \mathbb{R} \to \mathbb{R}$ be measurable and μ be any measure on $\mathbb{R}$. The function g defined by

$$g(s) = \operatorname*{ess\,sup}_{q \ge s} f(q)$$

is decreasing and lower-semicontinuous, where the essential supremum is with respect to μ.

Proof: The monotonicity is obvious. We only prove lower-semicontinuity, which, in particular, also implies right-continuity. Let $s_n \to s$. By definition of the essential supremum, for any δ, we have $\mu\{q \ge s : f(q) > g(s) - \delta\} > 0$. By the dominated convergence theorem, $\mu\{q \ge s_n : f(q) > g(s) - \delta\} \to \mu\{q \ge s : f(q) > g(s) - \delta\}$. Hence there exists N such that $\mu\{q \ge s_n : f(q) > g(s) - \delta\} > 0$ for all $n \ge N$, which implies that $g(s_n) \ge g(s) - \delta$ for all $n \ge N$. By the arbitrariness of δ, we have $\liminf_{n \to \infty} g(s_n) \ge g(s)$, completing the proof of the lower semi-continuity. □

Lemma 6. Under the conditions of Theorem 1, for any $u \ge 0$,

$$\begin{aligned} \lim_{n \to \infty} \frac{\log((1 - F_n(u\sqrt{2 \log n})) \wedge F_n(-u\sqrt{2 \log n}))}{\log n} \\ = \upsilon(u) \triangleq \operatorname*{ess\,sup}_{q \ge u} \{\alpha(q) - q^2\}. \end{aligned}$$

Proof: First assume that $u > 0$. Then

$$\begin{aligned} &1 - F_n(u\sqrt{2 \log n}) \\ &\quad = \int_{y \ge u\sqrt{2 \log n}} \exp(\ell_n(y)) \phi(y) \mathrm{d}y \\ &\quad = \sqrt{\frac{\log n}{\pi}} \int_{q \ge u} \exp(\ell_n(q\sqrt{2 \log n})) n^{-q^2} \mathrm{d}q \\ &\quad = n^{\upsilon(u) + o(1)}, \end{aligned}$$

where the last equality follows from Lemma 3. The proof for $u < 0$ is completely analogous. □

B. Proofs in Section III

Proof of Theorem 1: Let $W \sim \mathcal{N}(0, 1)$. Put $\nu_n = (1 - n^{-\beta})\mathcal{N}(0, 1) + n^{-\beta} G_n$. By the assumption that $G_n \ll \Phi$, we have $\nu_n \ll \mathcal{N}(0, 1)$. Denote the likelihood ratio by $L_n = \frac{g_n}{\varphi} = \exp(\ell_n)$. Then

$$\frac{\mathrm{d}\nu_n}{\mathrm{d}\Phi} = 1 + n^{-\beta}(\exp(\ell_n) - 1). \quad (74)$$

(Direct part) Recall the notation $\beta^\sharp$ defined in (28), which can be equivalently written as

$$\beta^\sharp = \frac{1}{2} + \operatorname*{ess\,sup}_{u \in \mathbb{R}} \left\{ \alpha_+(u) - u^2 + \frac{u^2 \wedge 1}{2} \right\}.$$

Assuming (29), we show that $\underline{\beta}^* \ge \beta^\sharp$ by lower bounding the Hellinger distance. To this end, fix an arbitrary $\delta > 0$.

Let $\beta = \beta^\sharp - 2\delta$. Denote by λ the Lebesgue measure on the real line. By definition of the essential supremum, $\lambda\{u : \alpha_+(u) - u^2 + \frac{u^2\wedge 1}{2} \geq \beta + \delta - \frac{1}{2}\} > 0$. Since $-u^2 + \frac{u^2\wedge 1}{2} \leq 0$ for all u and $\beta + \delta - \frac{1}{2} \geq -\delta$, we must have $\lambda\{u : \alpha(u) - u^2 + \frac{u^2\wedge 1}{2} \geq \beta + \delta - \frac{1}{2}, \alpha(u) \geq 0\} > 0$. Since, by assumption, $\lambda\{u : \alpha(u) > 0\} > 0$, there exists $0 < \epsilon \leq \frac{\delta}{2}$, such that

$$\lambda\left\{u : \alpha(u) - u^2 + \frac{u^2 \wedge 1}{2} \geq \beta + \delta - \frac{1}{2}, \alpha(u) \geq 2\epsilon\right\} > 0. \tag{75}$$

By assumption (29), there exists $N_\epsilon \in \mathbb{N}$ such that

$$\ell_n(u\sqrt{2\log n}) \geq (\alpha(u) - \epsilon)\log n \tag{76}$$

holds for all $u \in \mathbb{R}$ and all $n \geq N_\epsilon$. From (75), we have either

$$\lambda(E_1) > 0 \tag{77}$$

or

$$\lambda(E_2) > 0, \tag{78}$$

where

$$E_1 \triangleq \left\{u : |u| \leq 1, \alpha(u) - \frac{u^2}{2} \geq \beta + \delta - \frac{1}{2}, \alpha(u) \geq 2\epsilon\right\}$$
$$E_2 \triangleq \left\{u : |u| \geq 1, \alpha(u) - u^2 \geq \beta + \delta - 1, \alpha(u) \geq 2\epsilon\right\}.$$

Next we discuss these two cases separately.

Case I: Assume (77). Let

$$U = \frac{W}{\sqrt{2\log n}} \sim \mathcal{N}\left(0, \frac{1}{2\log n}\right). \tag{79}$$

The square Hellinger distance can be lower bounded as follows:

$$\begin{aligned}
&H_n^2(\beta)\\
&= H^2(P, \nu_n) = \int\left(\sqrt{\frac{\mathrm{d}\nu_n}{\mathrm{d}P}} - 1\right)^2 \mathrm{d}P\\
&= \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(\exp(\ell_n(U\sqrt{2\log n})) - 1)} - 1\right)^2\right] &(80)\\
&\geq \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(\exp(\ell_n(U\sqrt{2\log n})) - 1)} - 1\right)^2 \mathbf{1}_{\{U\in E_1\}}\right]\\
&\geq \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(n^{\alpha(U)-\epsilon} - 1)} - 1\right)^2 \mathbf{1}_{\{U\in E_1\}}\right] &(81)\\
&\geq \frac{(\sqrt{2}-1)^2}{4}\mathbb{E}\left[n^{(\alpha(U)-\epsilon-\beta)\wedge 2(\alpha(U)-\epsilon-\beta)}\mathbf{1}_{\{U\in E_1\}}\right] &(82)\\
&= \frac{(\sqrt{2}-1)^2\sqrt{\log n}}{4\sqrt{\pi}}\int_{E_1} n^{(\alpha(u)-\epsilon-\beta)\wedge 2(\alpha(u)-\epsilon-\beta)-u^2}\mathrm{d}u &(83)\\
&\geq \frac{(\sqrt{2}-1)^2\sqrt{\log n}}{4\sqrt{\pi}} n^{-1+\frac{\delta}{2}}\lambda(E_1) &(84)
\end{aligned}$$

where

- (80): By (74).
- (81): By Lemma 4.1 and (76).
- (82): Without loss of generality, we can assume that $n^\epsilon \geq 2$. Then applying the lower bound in Lemma 4.2 yields the desired inequality.
- (83): We used the density of U defined in (79).
- (84): Given that $|u| \leq 1$ and $\alpha(u) - u^2/2 \geq \beta + \delta - \frac{1}{2}$, we have both $\alpha(u) - \epsilon - \beta - u^2 \geq -\frac{1+v^2}{2} + \delta - \epsilon \geq -1 + \frac{\delta}{2}$ and $2\alpha(u) - 2\epsilon - 2\beta - u^2 \geq -1 + 2\delta - 2\epsilon \geq -1 + \delta$.

Case II: Now we assume (78). Following analogous steps as in the previous case, we have

$$\begin{aligned}
H_n^2(\beta) &\geq \frac{(\sqrt{2}-1)^2\sqrt{\log n}}{4\sqrt{\pi}}\int_{E_2} n^{(\alpha(u)-\epsilon-\beta)\wedge 2(\alpha(u)-\epsilon-\beta)-u^2}\\
&\geq \frac{(\sqrt{2}-1)^2\sqrt{\log n}}{4\sqrt{\pi}} n^{-1+\frac{\delta}{2}}\lambda(E_2) &(85)
\end{aligned}$$

where (85) is due to the following: Since $|u| \geq 1$ and $\alpha(u) - u^2 \geq \beta + \delta - 1$, we have both $\alpha(u) - \epsilon - \beta - u^2 \geq \delta - \epsilon - 1 \geq -1 + \frac{\delta}{2}$ and $2\alpha(u) - 2\epsilon - 2\beta - u^2 \geq v^2 - 2 + 2\delta - 2\epsilon \geq -1 + \delta$.

Combining (84) and (85) we conclude that $H_n^2(\beta) = \omega(n^{-1})$. By the arbitrariness of $\delta > 0$ and the alternative definition of $\underline{\beta}^*$ in (25), the proof of $\underline{\beta}^* \geq \beta^\sharp$ is completed.

(Converse part) Fix an arbitrary $\delta > 0$. Let

$$\beta = \beta^\sharp + 2\delta. \tag{86}$$

We upper bound the Hellinger integral as follows: First note that

$$\begin{aligned}
H_n^2(\beta) = \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(L_n - 1)} - 1\right)^2 \mathbf{1}_{\{L_n\geq 1\}}\right]\\
+\mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(L_n - 1)} - 1\right)^2 \mathbf{1}_{\{L_n\leq 1\}}\right]. &(87)
\end{aligned}$$

Applying Lemma 4.1, we have

$$\begin{aligned}
&\mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(L_n - 1)} - 1\right)^2 \mathbf{1}_{\{L_n\leq 1\}}\right]\\
&\qquad \leq (\sqrt{1 - n^{-\beta}} - 1)^2 \leq n^{-2\beta} = o(n^{-1}), &(88)
\end{aligned}$$

since $\beta > \beta^\sharp \geq \frac{1}{2}$ by (86). Consequently, the asymptotics of the Hellinger integral $H_n^2(\beta)$ is dominated by the first term in (87), denoted by a_n, which we analyze below using the Laplace method.

By (30), there exists $N_\delta \in \mathbb{N}$ such that

$$\ell_n(u\sqrt{2\log n}) \leq (\alpha(u) + \delta)\log n \tag{89}$$

holds for all $u \in \mathbb{R}$ and all $n \geq N_\delta$. Then

$$\begin{aligned}
a_n &\triangleq \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(L_n - 1)} - 1\right)^2 \mathbf{1}_{\{L_n\geq 1\}}\right]\\
&= \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(\exp(\ell_n(U\sqrt{2\log n})) - 1)} - 1\right)^2 \mathbf{1}_{\{L_n\geq 1\}}\right]\\
&\leq \mathbb{E}\left[\left(\sqrt{1 + n^{-\beta}(n^{\alpha(U)+\delta} - 1)} - 1\right)^2 \mathbf{1}_{\{\alpha(U)\geq -\delta\}}\right] &(90)\\
&\leq \mathbb{E}\left[\left(\sqrt{1 + n^{\alpha(U)+\delta-\beta}} - 1\right)^2\right]\\
&\leq \mathbb{E}\left[n^{(2(\alpha(U)+\delta-\beta))\wedge(\alpha(U)+\delta-\beta)}\right] &(91)\\
&= \sqrt{\frac{\log n}{\pi}}\int n^{(2(\alpha(u)+\delta-\beta))\wedge(\alpha(u)+\delta-\beta)-u^2}\mathrm{d}u &(92)
\end{aligned}$$

where (90) and (91) are due to (89) and Lemma 4.2, respectively. Next we apply Lemma 3 to analyze the exponent of (92). First we verify the integrability condition:

$$\begin{aligned}
\int n^{(2(\alpha(u)+\delta-\beta))\wedge(\alpha(u)+\delta-\beta)-u^2}\mathrm{d}u &\leq n^{\delta-\beta}\int n^{\alpha(u)-u^2}\mathrm{d}u\\
&< \infty,
\end{aligned}$$

in view of (32). Applying (71) to (92), we have

$$a_n \le n^{\operatorname{ess\,sup}_{u\in\mathbb{R}}\{(2(\alpha(u)+\delta-\beta))\wedge(\alpha(u)+\delta-\beta)-u^2\}+o(1)}. \quad (93)$$

By (86), $\alpha(u)-u^2+\frac{u^2\wedge 1}{2} \le \beta-\frac{1}{2}-2\delta$ holds a.e. Consequently, $\alpha(u)-u^2 \le \beta-1-2\delta$ holds for almost every $u \in (-\infty,-1]\cup[1,\infty)$ and $\alpha(u)-\frac{u^2}{2} \le \beta-\frac{1}{2}-2\delta$ holds for almost every $u\in[-1,1]$. These conditions immediately imply that

$$(2(\alpha(u)+\delta-\beta))\wedge(\alpha(u)+\delta-\beta)-u^2 \le -1+\delta \quad (94)$$

holds a.e. Assembling (87) and (93), we conclude that $H_n^2(\beta)=o(n^{-1})$. By the arbitrariness of $\delta>0$ and the alternative definition of $\overline{\beta}^*$ in (26), the proof of $\overline{\beta}^* \le \beta^\sharp$ is completed.

Proof of Theorem 2: In view of the proof of Theorem 1, the desired (33) readily follows from combining (84), (85), (88) and (93).

Proof of Lemma 2: Put

$$c(t) \triangleq \int \exp(t(\alpha(u)-u^2))\mathrm{d}u. \quad (95)$$

(Necessity) Since $\int g_n = 1$, we have $\int g_n(u\sqrt{\log n})\mathrm{d}u = (\log n)^{-\frac{1}{2}}$. By assumption, $g_n(u\sqrt{\log n}) = n^{\alpha(u)-u^2+o(1)}$ uniformly in u. Therefore $\int n^{\alpha(u)-u^2}\mathrm{d}u = n^{o(1)}$. Hence $c(\log n) < \infty$. It then follows from Lemma 3 that $\operatorname{ess\,sup}_u\{\alpha(u)-u^2\}=0$. Hence $\alpha(u)\le u^2$ a.e.

(Sufficiency) Let α be a measurable function satisfying (32). Then $c(t)<\infty$ for some $t>0$. By Hölder's inequality, $c(t)<\infty$ for all $t>0$. Let G_n be a probability measure with the density

$$g_n(y)=\frac{1}{c(\log n)\sqrt{\log n}}\exp\left\{\alpha\left(\frac{y}{\sqrt{2\log n}}\right)\log n-\frac{y^2}{2}\right\},$$

which is a legitimate density function in view of (95). By Lemma 3, $c(\log n)=n^{o(1)}$ as $n\to\infty$. Then the log-likelihood ratio satisfies $\ell_n(u\sqrt{\log n}) = \log\frac{\sqrt{2\pi}}{c(\log n)\sqrt{\log n}}+\alpha(u)$, which fulfills (31) uniformly.

For convolutional models, the convexity of α is inherited from the geometric properties of the log-likelihood ratio in the normal location model: Since $y\mapsto\log\frac{\mathbb{E}[\varphi(y-X)]}{\varphi(y)}$ is convex for any random variable X (see, e.g., [37, Property 3] and [38]), we have $\ell_n(((1-t)u+tv)\sqrt{2\log n}) \le (1-t)\ell_n(u\sqrt{2\log n})+t\ell_n(v\sqrt{2\log n})$ for any $t\in[0,1]$ and $u,v\in\mathbb{R}$. Dividing both sides by $\log n$ and sending $n\to\infty$, we have $\alpha((1-t)u+tv)\le(1-t)\alpha(u)+t\alpha(v)$.

Proof of Corollary 1: Since $g_n=\varphi * p_n$, we have

$$\begin{aligned} g_n(u\sqrt{2\log n}) &= \int_{\mathbb{R}}\varphi(u\sqrt{2\log n}-x)p_n(x)\mathrm{d}x \\ &= \sqrt{2\log n}\int_{\mathbb{R}}\varphi((u-t)\sqrt{2\log n})p_n(x\sqrt{2\log n})\mathrm{d}x \\ &= n^{o(1)}\int_{\mathbb{R}} n^{-(u-t)^2-f(t)+o(1)}\mathrm{d}x \\ &= n^{-\operatorname{ess\,inf}_{z\in\mathbb{R}}\{(u-t)^2+f(t)\}+o(1)} \end{aligned}$$

where the last equality follows from Lemma 3. Plugging the above asymptotics into $\ell_n=\log\frac{g_n}{\varphi}$, we see that (31) is fulfilled uniformly in $u\in\mathbb{R}$ with $\alpha(u)=u^2-\operatorname{ess\,inf}_{z\in\mathbb{R}}\{(u-\sqrt{r}z)^2+|z|^\tau\}$. Applying Theorem 1, we obtain

$$\begin{aligned} \beta^* &= \frac{1}{2}+\operatorname*{ess\,sup}_{u\in\mathbb{R}}\operatorname*{ess\,sup}_{t\in\mathbb{R}}\left\{-(u-t)^2-f(t)+\frac{u^2\wedge 1}{2}\right\} \\ &= \frac{1}{2}+\operatorname*{ess\,sup}_{t\in\mathbb{R}}\left\{-f(t)+\operatorname*{ess\,sup}_{u\in\mathbb{R}}\left\{-(u-t)^2+\frac{u^2\wedge 1}{2}\right\}\right\} \\ &= \sup_{t\in\mathbb{R}}\{\beta^*_{\mathrm{IDJ}}(t^2)-f(t)\} \end{aligned}$$

where the last step follows from the (53).

Proof of Theorem 3: Let $W_n\sim Q_n$. Put $\nu_n=(1-n^{-\beta})\Phi+n^{-\beta}G_n$. Since $G_n\ll Q_n$ by assumption, we also have $\nu_n\ll P$. Denote the likelihood ratio (Radon-Nikodym derivative) by $L_n=\frac{\mathrm{d}G_n}{\mathrm{d}Q_n}=\exp(\ell_n)$. Then

$$\frac{\mathrm{d}\nu_n}{\mathrm{d}Q_n}=1+n^{-\beta}(\exp(\ell_n)-1). \quad (96)$$

The proof proceeds analogously as in that of Theorem 1. Instead of introducing the random variable U in (79) for the Gaussian case, we apply the quantile transformation to generate the distribution of W_n: Let U be uniformly distributed on the unit interval. Then $S=\log\frac{1}{U}$ which is exponentially distributed. Putting $S_n=\frac{S}{\log n}$, we have

$$W_n\overset{(\mathrm{d})}{=}z_n(U)=z_n\left(n^{-S_n}\right)\overset{(\mathrm{d})}{=}z_n\left(1-n^{-S_n}\right). \quad (97)$$

Set $r_n(s)=\ell_n\circ z_n(n^{-s})$ and $t_n(s)=\ell_n\circ z_n(1-n^{-s})$, which satisfy

$$\sup_{s\ge\log_n 2}|r_n(s)-\alpha_0(s)\log n|\le\delta\log n \quad (98)$$

$$\sup_{s\ge\log_n 2}|t_n(s)-\alpha_1(s)\log n|\le\delta\log n \quad (99)$$

for all sufficiently large n. For the converse proof, similar to (87), we can write the square Hellinger distance as an expectation with respect to S_n:

$$\begin{aligned} &H_n^2(\beta) \\ &= \mathbb{E}\left[\left(\sqrt{1+n^{-\beta}(\exp(\ell_n(z_n(U)))-1)}-1\right)^2\mathbf{1}_{\left\{0<U<\frac{1}{2}\right\}}\right] \\ &\quad+\mathbb{E}\left[\left(\sqrt{1+n^{-\beta}(\exp(\ell_n(z_n(1-U)))-1)}-1\right)^2\mathbf{1}_{\left\{0<U\le\frac{1}{2}\right\}}\right] \\ &= \mathbb{E}\left[\left(\sqrt{1+n^{-\beta}(\exp(r_n(S_n))-1)}-1\right)^2\mathbf{1}_{\{S_n>\log_n 2\}}\right] \\ &\quad+\mathbb{E}\left[\left(\sqrt{1+n^{-\beta}(\exp(t_n(S_n))-1)}-1\right)^2\mathbf{1}_{\{S_n\ge\log_n 2\}}\right]. \end{aligned}$$

Define the events $F_1=\{S_n>\log_n 2, r_n(S_n)\ge 0\}$ and $F_2=\{S_n\ge\log_n 2, t_n(S_n)\ge 0\}$. Analogous to (88), by truncating the log-likelihood ratio at zero, we can show that the Hellinger

distance is dominated by the following:

$$\mathbb{E}\Big[\Big(\sqrt{1+n^{-\beta}(\exp(r_n(S_n))-1)}-1\Big)^2\mathbf{1}_{F_1}\Big] + \mathbb{E}\Big[\Big(\sqrt{1+n^{-\beta}(\exp(t_n(S_n))-1)}-1\Big)^2\mathbf{1}_{F_2}\Big] \quad (100)$$

$$\le \mathbb{E}\Big[\Big(\sqrt{1+n^{-\beta}(n^{\alpha_0(S_n)+\delta}-1)}-1\Big)^2 + \Big(\sqrt{1+n^{-\beta}(n^{\alpha_1(S_n)+\delta}-1)}-1\Big)^2\Big] \quad (101)$$

$$\le 2\,\mathbb{E}\Big[n^{2(\alpha_0\vee\alpha_1(U)+\delta-\beta)\wedge(\alpha_0\vee\alpha_1(U)+\delta-\beta)}\Big] \quad (102)$$

$$\le n^{-1-\delta} \quad (103)$$

where (101) from (98)–(99) and (103) from (92)–(94). The direct part of the proof is entirely analogous to that of Theorem 1 by lower bounding the integral in (100).

C. Proof of Theorem 4

Proof: Let $U_i = \Phi(X_i)$, which is uniformly distributed on [0, 1] under the null hypothesis. With a change of variable, we have

$$\mathrm{HC}_n = \sqrt{n}\sup_{t\in\mathbb{R}}\frac{|\mathbb{F}_n(t)-\Phi(t)|}{\sqrt{\Phi(t)\bar{\Phi}(t)}} \quad (104)$$

$$= \sqrt{n}\sup_{0<u<1}\frac{|\mathbb{F}_n(\Phi^{-1}(u))-u|}{\sqrt{u(1-u)}}, \quad (105)$$

which satisfies that $\frac{\mathrm{HC}_n}{\sqrt{2\log\log n}} \xrightarrow{\mathbb{P}} 1$ [30, p. 604]. Therefore the Type-I error probability of the test (50) vanishes for any choice of $\delta > 0$. It remains to show that $\mathrm{HC}_n = \omega_{\mathbb{P}}(\log\log n)$ under the alternative. To this end, fix $0 < s < 1$ and put $r_{n,s} = \Phi(\sqrt{2s\log n})$ and $\rho_{n,s} = (1-n^{-\beta})\Phi(\sqrt{2s\log n}) + n^{-\beta}G_n(\sqrt{2s\log n})$. By (104), we have

$$\mathrm{HC}_n \ge V_n(s) \triangleq \sqrt{n}\frac{\mathbb{F}_n(\sqrt{2s\log n})-r_{n,s}}{\sqrt{r_{n,s}(1-r_{n,s})}} \quad (106)$$

$$= \frac{N_n(s)-nr_{n,s}}{\sqrt{nr_{n,s}(1-r_{n,s})}}, \quad (107)$$

where $N_n(s) \triangleq \sum_{i=1}^n \mathbf{1}_{\{X_i\ge\sqrt{2s\log n}\}}$ is binomially distributed with sample size n and success probability $\rho_{n,s}$. Therefore

$$\mathbb{E}[V_n(s)] = \sqrt{n}\frac{\rho_{n,s}-r_{n,s}}{\sqrt{r_{n,s}(1-r_{n,s})}} = n^{\frac{1}{2}-\beta}\frac{G_n(\sqrt{2s\log n})-r_{n,s}}{\sqrt{r_{n,s}(1-r_{n,s})}}. \quad (108)$$

and

$$\mathsf{var}V_n(s) = \frac{\rho_{n,s}(1-\rho_{n,s})}{r_{n,s}(1-r_{n,s})}. \quad (109)$$

By Chebyshev's inequality,

$$\mathbb{P}\left\{V_n(s) \le \frac{1}{2}\mathbb{E}[V_n(s)]\right\} \le \frac{4\,\mathsf{var}V_n(s)}{\mathbb{E}[V_n(s)]^2} = \frac{4\rho_{n,s}(1-\rho_{n,s})}{n(\rho_{n,s}-r_{n,s})^2}.$$

By Lemma 6,

$$1-G_n(\sqrt{2s\log n}) = n^{\upsilon(s)+o(1)}, \quad (110)$$

where $\upsilon(s) = \operatorname{ess\,sup}_{q\ge s}\{\alpha(q)-q\} \ge -s$. Plugging (110) into (108) and (109) yields

$$\mathbb{E}[V_n(s)] = n^{\frac{1+s}{2}-\beta+\upsilon(s)+o(1)} \quad (111)$$

and

$$\mathbb{P}\{V\}_n(s) \le \frac{1}{2}\mathbb{E}[V_n(s)] \le n^{2\beta-s-1-2\upsilon(s)+o(1)} + n^{\beta-1-\upsilon(s)+o(1)}. \quad (112)$$

Suppose that $\beta < \frac{1+s}{2} + \upsilon(s)$. Then $\mathbb{E}[V_n(s)] = \omega(\sqrt{\log\log n})$. Moreover, we have $2\beta - s - 1 - 2\upsilon(s) < 0$ and $\beta - 1 - \upsilon(s) \le \frac{s-1}{2} < 0$ since $s < 1$. Combining (106), (111) and (112), we obtain

$$\mathbb{P}\left\{\mathrm{HC}_n > \sqrt{(2+\delta)\log\log n}\right\} = 1 - o(1),$$

that is, the Type-II error probability also vanishes. Consequently, a sufficient condition for the higher criticism test to succeed is

$$\beta < \sup_{0<s<1}\frac{1+s}{2} + \upsilon(s) \quad (113)$$

$$= \operatorname*{ess\,sup}_{0<s<1}\frac{1+s}{2} + \upsilon(s), \quad (114)$$

where (114) follows from the following reasoning: By [39, Proposition 3.5], the supremum and the essential supremum (with respect to the Lebesgue measure) coincide for all lower semi-continuous functions. Indeed, υ is lower semi-continuous by Lemma 5, and so is $s \mapsto \frac{1+s}{2} + \upsilon(s)$.

It remains to show that the right-hand side of (114) coincides with the expression of β^* in Theorem 1. Indeed, we have

$$\begin{aligned}\operatorname*{ess\,sup}_{0\le s\le 1}\{s+2\upsilon(s)\} &= \operatorname*{ess\,sup}_{0\le s\le 1}\left\{s+2\operatorname*{ess\,sup}_{q\ge s}\{\alpha(q)-q\}\right\}\\ &= \operatorname*{ess\,sup}_{q\ge 0}\operatorname*{ess\,sup}_{q\wedge 1\le s\le 1}\{2\alpha(q)-2q+s\}\\ &= \operatorname*{ess\,sup}_{q\ge 0}\{2\alpha(q)-2q+q\wedge 1\}.\end{aligned}$$

Here the second equality follows from interchanging the essential suprema: For any bi-measurable function $(x, y) \mapsto f(x, y)$,

$$\operatorname*{ess\,sup}_x\operatorname*{ess\,sup}_y f(x,y) = \operatorname*{ess\,sup}_y\operatorname*{ess\,sup}_x f(x,y) = \operatorname*{ess\,sup}_{x,y} f(x,y),$$

where the last essential supremum is with respect to the product measure. Thus the proof of the theorem is completed. □

APPENDIX

This appendix collects a few properties of total variation and Hellinger distances for mixture distributions.

Lemma 7. Let $0 \le \epsilon \le 1$ and $Q_1 \perp P$. Then

$$H^2(P,(1-\epsilon)Q_0+\epsilon Q_1) = 2(1-\sqrt{1-\epsilon}) + \sqrt{1-\epsilon}\,H^2(P,Q_0) \quad (115)$$

which satisfies

$$\frac{1}{4} \le \frac{H^2(P,(1-\epsilon)Q_0+\epsilon Q_1)}{\epsilon \vee H^2(P,Q_0)} \le 4 \quad (116)$$

Proof: Since $Q_1 \perp P$, there exists a measurable set E such that $P(E) = 0$ and $Q_1(E) = 1$. Then

$$\begin{aligned} H^2(P, (1-\epsilon)Q_0 + \epsilon Q_1) &= 2 - 2\int \sqrt{\mathrm{d}P((1-\epsilon)\mathrm{d}Q_0 + \epsilon \mathrm{d}Q_1)} \\ &= 2 - 2\sqrt{1-\epsilon}\int_{\circ E} \sqrt{\mathrm{d}P \mathrm{d}Q_0} \\ &= 2 - \sqrt{1-\epsilon}\,(2 - H^2(P, Q_0)). \end{aligned}$$

The inequalities in (116) follow from (115) and the facts that $\frac{\epsilon}{2} \le \sqrt{1-\epsilon} \le \epsilon$ and $0 \le H^2 \le 2$. □

Lemma 8. For any probability measures (P, Q), $\epsilon \mapsto H^2(P, (1-\epsilon)P + \epsilon Q)$ is decreasing on $[0, 1]$.

Proof: Fix $0 \le \epsilon < \epsilon' \le 1$. Since $(1-\epsilon)P + \epsilon Q = ((1-\epsilon')P + \epsilon' Q)\frac{\epsilon}{\epsilon'} + \frac{\hat{\epsilon}-\epsilon}{\epsilon'}P$, the convexity of $H^2(P, \cdot)$ yields

$$H^2((1-\epsilon)P + \epsilon Q, P) \le \frac{\epsilon}{\epsilon'} H^2((1-\epsilon')P + \epsilon' Q, P). \quad \square$$

We conclude this appendix by proving Lemma 1 presented in Section II-A.

Proof: By Lemma 8, the function $\beta \mapsto H_n^2(\beta)$ is decreasing, which, in view of the characterization (25)–(26), implies that $\underline{\beta}^* \le \overline{\beta}^*$. Thus it only remains to establish the rightmost inequality in (19). To this end, we show that as soon as β exceeds 1, $V_n(\beta)$ becomes $o(1)$ regardless of the choice of $\{G_n\}$: Fix $\beta > 1$. Then

$$\begin{aligned} V_n(\beta) &= \mathsf{TV}(\Phi^n, ((1-n^{-\beta})\Phi + n^{-\beta}G_n)^n) \\ &\le \mathsf{TV}(\delta_0^n, ((1-n^{-\beta})\delta_0 + n^{-\beta}\delta_1)^n) \qquad (117) \\ &= 1 - (1-n^{-\beta})^n \\ &\le n^{1-\beta} \\ &= o(1), \end{aligned}$$

where (117) follows from the data-processing inequality, which is satisfied for all f-divergences [40], in particular, the total variation: $\mathsf{TV}(P_Y, Q_Y) \le \mathsf{TV}(P_X, Q_X)$, where $Q_{Y|X} = P_{Y|X}$ is any probability transition kernel. □

Remark 4. While Lemma 8 is sufficient for our purpose in proving Lemma 1, it is unclear whether the monotonicity carries over to $\epsilon \mapsto \mathsf{TV}(P^n, ((1-\epsilon)P + \epsilon Q)^n)$, since product measures do not form a convex set. It is however easy to see that $\epsilon \mapsto \mathsf{TV}(P^n, ((1-\epsilon)P + \epsilon Q)^n)$ is decreasing, which follows from the proof of Lemma 8 with H^2 replaced by TV. It is also clear that $\epsilon \mapsto H^2(P^n, ((1-\epsilon)P + \epsilon Q)^n)$ is decreasing in view of (23).

In this appendix we show that (39) implies that $\beta^* = 1$, i.e., for any $\beta < 1$, the hypotheses in (11) can be tested reliably. Without loss of generality, we assume that $u \ge 1$. Then

$$\tau_n \triangleq G_n((\sqrt{2\log n}, \infty)) = n^{-o(1)},$$

We show that the total variation distance between the product measures converge to one. Put $A_n = (-\infty, \sqrt{2s\log n}]^n$. In view of the first inequality in (14), the total variation distance can be lower bounded as follows:

$$V_n(\beta) \ge \Phi^n(A_n) - ((1-n^{-\beta})\Phi + n^{-\beta}G_n)^n(A_n).$$

Using (44), we have

$$\Phi^n(A_n) = (1 - \bar{\Phi}(\sqrt{2s\log n}))^n = 1 - \frac{n^{1-s}}{\sqrt{4\pi s\log n}}(1 + o(1)).$$

On the other hand,

$$\begin{aligned} &((1-n^{-\beta})\Phi + n^{-\beta}G_n)^n(A_n) \\ &\quad = (1 - (1-n^{-\beta})\bar{\Phi}(\sqrt{2s\log n}) - n^{-\beta}\tau_n)^n \\ &\quad = (1 - n^{-s+o(1)} - n^{-\beta-s+o(1)} - n^{-\beta+o(1)})^n \\ &\quad = o(1) \end{aligned}$$

where the last equality is due to $0 < \beta < 1 \le s$. Therefore $V_n(\beta) = 1 - o(1)$ for any $\beta < 1$, which proves that $\beta^* = 1$.

In fact, the above derivation also shows that the following *maximum test* achieves vanishing probability of error: declare H_1 if and only if $\max_i |X_i| > |u|\sqrt{2\log n}$. In general the maximum test is suboptimal. For example, in the classical setting (2) where $G_n = \delta_{\mu_n}$, [8, Theorem 1.3] shows that the maximum test does not attain the Ingster-Donoho-Jin detection boundary for $\beta \in [\frac{1}{2}, \frac{3}{4}]$.

ACKNOWLEDGMENT

The paper has benefited from thorough suggestions by the Associate Editor and the anonymous reviewers.

REFERENCES

[1] R. L. Dobrushin, "A statistical problem arising in the theory of detection of signals in the presence of noise in a multi-channel system and leading to stable distribution laws," *Theory Probab. Appl.*, vol. 3, no. 2, pp. 161–173, 1958.

[2] X. J. Jeng, T. T. Cai, and H. Li, "Optimal sparse segment identification with application in copy number variation analysis," *J. Amer. Statist. Assoc.*, vol. 105, no. 491, pp. 1156–1166, 2010.

[3] L. Cayon, J. Jin, and A. Treaster, "Higher criticism statistic: Detecting and identifying non-Gaussianity in the WMAP first-year data," *Monthly Notices R. Astron. Soc.*, vol. 362, no. 3, pp. 826–832, 2005.

[4] J. Jin, J. L. Starck, D. L. Donoho, N. Aghanim, and O. Forni, "Cosmological non-Gaussian signature detection: Comparing performance of different statistical tests," *EURASIP J. Appl. Signal Process.*, vol. 15, pp. 2470–2485, Jan. 2005.

[5] M. Kulldorff, R. Heffernan, J. Hartman, R. Assunção, and F. Mostashari, "A space–time permutation scan statistic for disease outbreak detection," *PLoS Med.*, vol. 2, no. 3, p. e59, 2005.

[6] T. T. Cai, J. X. Jeng, and J. Jin, "Optimal detection of heterogeneous and heteroscedastic mixtures," *J. R. Statist. Soc. Ser. B (Methodological)*, vol. 73, no. 5, pp. 629–662, Nov. 2011.

[7] T. T. Cai, J. Jin, and M. G. Low, "Estimation and confidence sets for sparse normal mixtures," *Ann. Statist.*, vol. 35, no. 6, pp. 2421–2449, 2007.

[8] D. L. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist.*, vol. 32, no. 3, pp. 962–994, 2004.

[9] P. Hall, Y. Pittelkow, and M. Ghosh, "Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes," *J. R. Statist. Soc. Ser. B (Methodological)*, vol. 70, no. 1, pp. 159–173, 2008.

[10] M. V. Burnashev and I. A. Begmatov, "On a problem of detecting a signal that leads to stable distributions," *Theory Probab. Appl.*, vol. 35, no. 3, pp. 556–560, 1990.

[11] Y. I. Ingster, "On some problems of hypothesis testing leading to infinitely divisible distributions," *Math. Methods Statist.*, vol. 6, no. 1, pp. 47–69, 1997.

[12] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *Ann. Math. Statist.*, vol. 36, no. 2, pp. 369–401, 1965.

[13] O. Zeitouni, J. Ziv, and N. Merhav, "When is the generalized likelihood ratio test optimal," *IEEE Trans. Inf. Theory*, vol. 38, no. 5, pp. 1597–1602, Sep. 1992.

[14] J. Unnikrishnan, D. Huang, S. P. Meyn, A. Surana, and V. V. Veeravalli, "Universal and composite hypothesis testing via mismatched divergence," *IEEE Trans. Inf. Theory*, vol. 57, no. 3, pp. 1587–1603, Mar. 2011.
[15] Y. I. Ingster, "Adaptive detection of a signal of growing dimension. I," *Math. Methods Statist.*, vol. 10, no. 4, pp. 395–421, 2001.
[16] Y. I. Ingster, "Adaptive detection of a signal of growing dimension. II," *Math. Methods Statist.*, vol. 11, no. 1, pp. 37–68, 2002.
[17] L. Jager and J. A. Wellner, "Goodness-of-fit tests via phi-divergences," *Ann. Statist.*, vol. 35, no. 5, pp. 2018–2053, 2007.
[18] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. Math. Statist. Probab.*, 1961, pp. 547–561.
[19] P. Hall and J. Jin, "Innovated higher criticism for detecting sparse signals in correlated noise," *Ann. Statist.*, vol. 38, no. 3, pp. 1686–1732, 2010.
[20] E. Arias-Castro, D. L. Donoho, and X. Huo, "Near-optimal detection of geometric objects by fast multiscale methods," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2402–2425, Jul. 2005.
[21] E. Arias-Castro, E. J. Candés, H. Helgason, and O. Zeitouni, "Searching for a trail of evidence in a maze," *Ann. Statist.*, vol. 36, no. 4, pp. 1726–1757, 2008.
[22] E. Arias-Castro, E. J. Candés, and Y. Plan, "Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism," *Ann. Statist.*, vol. 39, no. 5, pp. 2533–2556, 2011.
[23] Y. I. Ingster and I. A. Suslina, *Nonparametric goodness-of-Fit testing under Gaussian Models*. New York, NY, USA: Springer-Verlag, 2003.
[24] E. Arias-Castro and M. Wang, "Distribution-free tests for sparse heterogeneous mixtures," 2013.
[25] J. Neyman and E. S. Pearson, "On the problem of the most efficient tests of statistical hypotheses," *Phil. Trans. R. Soc. London Series A, Containing Papers Math. Phys. Character*, vol. 231, pp. 289–337, Jan. 1933.
[26] L. Le Cam, *Asymptotic Methods in Statistical Decision Theory*. New York, NY, USA: Springer-Verlag, 1986.
[27] L. de Haan and A. Ferreira, *Extreme Value Theory: An Introduction*. New York, NY, USA: Springer-Verlag, 2006.
[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. New York, NY, USA: Wiley, 1984.
[29] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. Boca Raton, FL, USA: CRC Press, 1992.
[30] G. R. Shorack and J. A. Wellner, *Empirical Processes with Applications to Statistics*. New York, NY, USA: Wiley, 1986.
[31] T. Taguchi, "On a generalization of Gaussian distribution," *Ann. Inst. Statist. Math., A*, vol. 30, no. 1, pp. 211–242, 1978.
[32] C. Butucea and Y. I. Ingster, "Detection of a sparse submatrix of a high-dimensional noisy matrix," *Bernoulli*, vol. 19, no. 5B, pp. 2652–2688, Nov. 2013.
[33] R. R. Bahadur, "Rates of convergence of estimates and test statistics," *Ann. Math. Statist.*, vol. 38, no. 2, pp. 303–324, 1967.
[34] J. L. Hodges and E. L. Lehmann, "The efficiency of some nonparametric competitors of the t-test," *Ann. Math. Statist.*, vol. 27, no. 2, pp. 324–335, 1956.
[35] H. S. Wieand, "A condition under which the Pitman and Bahadur approaches to efficiency coincide," *Ann. Statist.*, vol. 4, no. 5, pp. 1003–1011, 1976.
[36] A. Erdélyi, *Asymptotic Expansions*. New York, NY, USA: Dover, 1956.
[37] C. Hatsell and L. Nolte, "Some geometric properties of the likelihood ratio," *IEEE Trans. Inf. Theory*, vol. 17, no. 5, pp. 616–618, Sep. 1971.
[38] R. Esposito, "On a relation between detection and estimation in decision theory," *Inf. Control*, vol. 12, no. 2, pp. 116–120, 1968.
[39] H. X. Phu and A. Hoffmann, "Essential supremum and supremum of summable functions," *Numer. Funct. Anal. Optim.*, vol. 17, nos. 1–2, pp. 161–180, 1996.
[40] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, Jan. 1967.

Tony T. Cai received the Ph.D. degree from Cornell University, Ithaca, NY, in 1996. His research interests include high-dimensional inference, large-scale multiple testing, nonparametric function estimation, functional data analysis and statistical decision theory. He is the Dorothy Silberberg Professor of Statistics at the Wharton School of the University of Pennsylvania, Philadelphia, PA, USA. Dr. Cai is the recipient of the 2008 COPSS Presidents Award and a fellow of the Institute of Mathematical Statistics. He is the past editor of the Annals of Statistics.

Yihong Wu (S'10–M'13) received the B.E. degree from Tsinghua University, Beijing, China, in 2006 and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, in 2008 and 2011, all in electrical engineering. Before joining the Department of Electrical and Computer Engineering at University of Illinois at Urbana-Champaign as an assistant professor in 2013, he was a postdoctoral fellow with the Statistics Department, The Wharton School, University of Pennsylvania, Philadelphia, PA. His research interests are in information theory, high-dimensional statistics, decision theory, approximation theory and fractal geometry.

Dr. Wu was a recipient of the Princeton University Wallace Memorial Honorific Fellowship in 2010. He was a recipient of 2011 Marconi Society Paul Baran Young Scholar Award and the Best Student Paper Award at the 2011 IEEE International Symposiums on Information Theory (ISIT).