Detection of Sparse Mixtures: The Finite Alphabet Case

Jonathan G. Ligo*, George V. Moustakides †, Venugopal V. Veeravalli*
*University of Illinois at Urbana-Champaign
Urbana, IL 61801
†University of Patras and Rutgers University
Rio, GR 26500 and New Brunswick, NJ 08910

***Abstract*—We study the problem of testing between a sparse signal in noise, modeled as a mixture distribution, versus pure noise, with finite alphabet observations. We study the consistency and adaptivity of the tests as the mixture proportion tends to zero with number of observations. The finite alphabet assumption allows for application to inherently categorical data, where no useful ordering relationship on the alphabet typically exists. We construct and analyze a divergence-based adaptive test for finite alphabets and validate it on a quantized Gaussian signal detection problem.**

***Index Terms*—Detection theory, large deviations, error exponents, sparse detection, likelihood ratio test**

I. Introduction

We consider the problem of detecting an unknown sparse signal in noise, modeled as a mixture, where the unknown sparsity level decreases as the number of samples collected increases. Of particular interest is the case where the signal strength relative to the noise power is very small. This problem has mostly been studied in the case of standard Gaussian noise with a Gaussian signal [1]–[3] along with some extensions for non-Gaussian, but real-valued, signal and noise models [4]–[6]. The central results in the literature consist of conditions on the signal and noise such that the detection problem can be solved with vanishing error probability (or impossibility, thereof) and construction of *adaptive* tests that can be used to detect the unknown signal with only knowledge of the noise statistics. Applications include covert communications [3], [7]–[9], computational biology [10], [11], astrophysics [12] and machine learning [13].

In this paper, we focus on signals defined on a finite set (alphabet). The finite alphabet assumption allows for application to categorical data, which often occurs as features in machine learning or symbols in a communications constellation. These signals typically do not possess an ordering, so a straightforward application of real-valued techniques is not always sensible. Finite alphabets also arise from quantizing data from a larger (possibly uncountable) alphabet for reduced storage, communication and/or computational complexity. The quantization of a real-valued signal will be considered further in Sec. V, where we see quantizer designs can have a large effect on detector performance.

This work was supported in part by the US National Science Foundation under the grant CCF 1514245 through the University of Illinois at Urbana-Champaign, and under the Grant CIF 1513373, through Rutgers University.

Our contributions are conditions for when the detection problem is impossible and a characterization the rate of decay of the false alarm and miss detection probabilities for an oracle test when the problem is possible in Sec. III. In contrast to the problem of testing between n i.i.d. samples from two fixed distributions, where the error probabilities behave as e^{-cn} where c is determined by the Kullback-Leibler divergence between the hypotheses [14], [15], the finite mixture detection problem exhibits subexponential decay. We complement the oracle test analysis with theoretical and numerical analysis of a simple adaptive test with competitive error performance (Sec. IV and VI).

II. Problem Setup

Let $\{f_{0,n}\}, \{f_{1,n}\}$ be sequences of probability mass functions (PMFs) on the finite alphabet $\mathcal{X}$.

We consider the following sequence of composite hypothesis testing problems with sample size n, called the *finite (sparse) mixture detection problem*:

$$\mathsf{H}_{0,n} : X_1, \ldots, X_n \sim f_{0,n} \text{ i.i.d. (null)} \tag{1}$$

$$\mathsf{H}_{1,n} : X_1, \ldots, X_n \sim (1-\epsilon_n) f_{0,n} + \epsilon_n f_{1,n} \text{ i.i.d. (alternative)} \tag{2}$$

where $\{f_{0,n}\}$ is known with full support and $\{f_{1,n}\}$ is from some known family of sequences of PMFs $\mathcal{F}$ and $\{\epsilon_n\}$ is a sequence of positive numbers (called the *sparsity level*) such that $\epsilon_n \to 0$. We will also assume $n\epsilon_n \to \infty$ so that a typical realization of the alternative is distinguishable from the null.

Let $\mathsf{P}_{0,n}, \mathsf{P}_{1,n}$ denote the probability measure under $\mathsf{H}_{0,n}, \mathsf{H}_{1,n}$ respectively, and let $\mathsf{E}_{0,n}, \mathsf{E}_{1,n}$ be the corresponding expectations with respect to the particular $\{f_{0,n}\}$, $\{f_{1,n}\}$ and $\{\epsilon_n\}$. When convenient, we will drop the subscript n. We can think of $f_{1,n}$ as a signal distribution and $f_{0,n}$ as a noise distribution. Thus, we are concerned with the problem of detecting a sparse signal with on average a fraction of ϵ_n components of signal and the remaining components as noise.

Let $\mathsf{L}_n \triangleq \frac{f_{1,n}}{f_{0,n}}$ be the likelihood ratio between the signal and noise.

A hypothesis test δ_n between $\mathsf{H}_{0,n}$ and $\mathsf{H}_{1,n}$ is a function $\delta_n : (x_1, \ldots, x_n) \to \{0, 1\}$. We define the *probability of false alarm* for a hypothesis test δ_n between $\mathsf{H}_{0,n}$ and $\mathsf{H}_{1,n}$ as $\mathsf{P}_{\mathrm{FA}}(n) \triangleq \mathsf{P}_{0,n}(\delta_n = 1)$ and the *probability of missed detection* as $\mathsf{P}_{\mathrm{MD}}(n) \triangleq \mathsf{P}_{1,n}(\delta_n = 0)$.

A sequence of hypothesis tests $\{\delta_n\}$ is *consistent* if $\mathsf{P}_{\mathrm{FA}}(n)$, $\mathsf{P}_{\mathrm{MD}}(n) \to 0$ as $n \to \infty$. We say we have a *rate characterization* for a sequence of consistent hypothesis tests $\{\delta_n\}$ if we can write

$$\lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{FA}}(n)}{g_0(n)} = -c, \lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{MD}}(n)}{g_1(n)} = -d \quad (3)$$

where $g_0(n), g_1(n) \to \infty$ as $n \to \infty$ and $0 < c, d < \infty$. The rate characterization describes decay of the error probabilities for large sample sizes. All logarithms are natural. For the problem of testing between i.i.d. samples from two fixed distributions, the rate characterization has $g_0(n) = g_1(n) = n$ and c, d are called the *error exponents* [14]. In the mixture detection problem, g_0 and g_1 will be sublinear functions of n.

The log-likelihood ratio between $\mathsf{H}_{1,n}$ and $\mathsf{H}_{0,n}$ is

$$\mathrm{LLR}(n) = \sum_{i=1}^{n} \log(1 - \epsilon_n + \epsilon_n \mathsf{L}_n(X_i)). \quad (4)$$

In order to perform an *oracle rate* characterization for the mixture detection problem, we consider the sequence of oracle likelihood ratio tests (LRTs) between $\mathsf{H}_{0,n}$ and $\mathsf{H}_{1,n}$ (i.e. with $\epsilon_n, f_{0,n}, f_{1,n}$ known):

$$\delta_n(X_1, \ldots, X_n) \triangleq \begin{cases} 1 & \mathrm{LLR}(n) \geq 0 \\ 0 & o.w. \end{cases}. \quad (5)$$

It is well known that (5) is optimal for testing between $\mathsf{H}_{0,n}$ and $\mathsf{H}_{1,n}$ in the sense of minimizing $\frac{\mathsf{P}_{\mathrm{FA}}(n)+\mathsf{P}_{\mathrm{MD}}(n)}{2}$, which is the average probability of error when the null and alternative are assumed to be equally likely [14]. It is valuable to analyze P_{FA} and P_{MD} separately since many applications incur different penalties for false alarms and missed detections.

III. Oracle Rate Analysis

In this section, we analyze the error probabilities of the Likelihood Ratio Test given by (5).

We first define some notation. A sequence a_n is $= O(b_n)$ if $\limsup_{n\to\infty} |\frac{a_n}{b_n}| \leq C$ for some constant $C \geq 0$. If $C = 0$, then $a_n = o(b_n)$. If $a_n = O(b_n)$ then $b_n = \Omega(a_n)$. If $a_n = o(b_n)$, then $b_n = \omega(a_n)$. If $a_n = O(b_n)$ and $a_n = \Omega(b_n)$, then $a_n = \Theta(b_n)$. We use $\mathsf{E}[\cdot; S] = \mathsf{E}[\cdot \mathbb{1}_S]$ as a convenient shorthand.

For the purposes of presentation, we assume the alphabet $\mathcal{X}$ can be partitioned into sets $\mathcal{X}_0, \mathcal{X}_1, \mathcal{X}_\infty$ where

$$\mathcal{X}_0 = \{x \in \mathcal{X} : \epsilon_n \mathsf{L}_n(x) = o(1)\}$$

$$\mathcal{X}_1 = \{x \in \mathcal{X} : \epsilon_n \mathsf{L}_n(x) = \Theta(1)\}$$

$$\mathcal{X}_\infty = \{x \in \mathcal{X} : \epsilon_n \mathsf{L}_n(x) = \omega(1)\}.$$

This partitioning is sufficiently general to include almost all cases of interest. Note $\mathsf{P}_{0,n}[\mathcal{X}_0] \to 1$. By inspecting the LRT test statistic (4), for sufficiently large n, we see that samples from $\mathcal{X}_1, \mathcal{X}_\infty$ always contribute positively the LLR, whereas samples from $\mathcal{X}_0$ may provide positive or negative contributions to the LLR depending on if $\epsilon_n(\mathsf{L}_n - 1) > 0$ or otherwise. We will also assume that for $x \in \mathcal{X}_\infty$, we have $\epsilon_n \mathsf{L}_n(x) = O(n^c)$ and $\epsilon_n \mathsf{L}_n(x) = \omega(n^d)$ for some $c, d > 0$, i.e., the likelihood ratio between the signal distribution and noise distribution grows polynomially. This assumption prevents degenerate behavior in Thm 3.1 and Thm 3.2, such as $f_{0,n}(x) = e^{-2n}$ and $f_{1,n}(x) = e^{-n}$ for some $x \in \mathcal{X}_\infty$.

Our first result is for "weak signals", where the error behavior is determined by the behavior of the hypotheses on $\mathcal{X}_0$.

Theorem 3.1: Assume that $\epsilon_n \mathsf{P}_{f_1}[\mathcal{X}_1 \cup \mathcal{X}_\infty] = o(\epsilon_n^2 D_n^2)$ where $D_n^2 = \mathsf{E}_0[(\mathsf{L}_n - 1)^2; \mathcal{X}_0]$. Also, assume $n\epsilon_n^2 D_n^2 \to \infty$ and

$$\frac{\max_{x\in\mathcal{X}_\infty} \log^2(1 + \epsilon_n(L_n - 1))}{n\epsilon_n^2 D_n^2} \to 0. \quad (6)$$

Then,

$$\lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2 D_n^2} = \lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{MD}}(n)}{n\epsilon_n^2 D_n^2} = -\frac{1}{8}. \quad (7)$$

If (6) is violated, the equalities and limits in (7) can be replaced by $\leq$ and $\limsup$, respectively.

Proof: We sketch the argument for P_{FA}. The results for P_{MD} follow by a change of measure to the null distribution. When $\mathcal{X} = \mathcal{X}_0$, this result is a direct application of Theorem 3.1 from [6]. Let $\Lambda_n(s) = E_0[(1 + \epsilon_n(\mathsf{L}_n(X_1) - 1))^s]$. A Chernoff bound furnishes $\mathsf{P}_{\mathrm{FA}}(n) \leq \Lambda_n(s)^n$ for any $s \in [0, 1]$. Calculating $\Lambda_n(s)$ by applying a Maclaurin series for $\mathcal{X}_0$ and changing measure to $f_{1,n}$ by multiplying and dividing the integrand by $\epsilon_n \mathsf{L}_n$ on $\mathcal{X}_1 \cup \mathcal{X}_\infty$ shows that $\Lambda_n(s) = 1 - \frac{s(1-s)}{2}\epsilon_n^2 D_n^2(1 + o(1))$. Choosing $s = 1/2$ establishes an upper bound on the rate. The corresponding lower bound is established by changing measure to the tilted distribution

$$\tilde{f}(x) = \frac{(1 + \epsilon_n(L_n(x) - 1))^{s_n}}{\Lambda_n(s_n)} f_0(x) \quad (8)$$

where $s_n = \arg\min_{s\in[0,1]} \Lambda_n(s)$. Under the assumptions of the theorem, it can be shown that $\log(1 + \epsilon_n(\mathsf{L}_n - 1))$ has mean zero and variance $\sigma_n^2 = \Theta(\epsilon_n^2 D_n^2)$, and the Lindeberg-Feller Central Limit Theorem shows that $\frac{\mathrm{LLR}(n)}{\sqrt{n}\sigma_n}$ converges to a standard normal distribution. Proceeding similarly to Cramer's theorem (Thm I.4, [16]) or the lower bound in Theorem 3.1 from [6] establishes the lower bound. ∎

The condition (6) is automatically satisfied if $\mathcal{X}_\infty = \emptyset$ or $\epsilon_n^2 D_n^2 = \omega\left(\frac{\log^2 n}{n}\right)$. Note that even in the absence of this condition, our rate guarantee holds modulo a small polylogarithmic backoff from the detection limit given in Thm. 3.3.

Our next result is for "strong signals", where the error behavior is determined by the behavior of the hypotheses on $\mathcal{X}_1 \cup \mathcal{X}_\infty$.

Theorem 3.2: Assume that $\epsilon_n \mathsf{P}_{f_1}[\mathcal{X}_1 \cup \mathcal{X}_\infty] = \omega(\epsilon_n^2 D_n^2)$ and $n\epsilon_n \mathsf{P}_{f_1}[\mathcal{X}_1 \cup \mathcal{X}_\infty] \to \infty$.

Also, assume that

$$\frac{\max_{x\in\mathcal{X}_\infty}\log^2(1+\epsilon_n(\mathsf{L}_n-1))}{n\epsilon_n^2D_n^2+n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1]+n\log n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_\infty]}\to 0 \quad (9)$$

Then,

$$\frac{\log\mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1\cup\mathcal{X}_\infty]},\frac{\log\mathsf{P}_{\mathrm{MD}}(n)}{n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1\cup\mathcal{X}_\infty]}=-\Theta(1). \quad (10)$$

where $\Theta(1)$ denotes some quantity upper and lower bounded by positive constants. If (9) is violated, then equalities in (10) can be replaced with $\leq$ signs and $\Theta(1)$ a positive constant.

Proof: We sketch the argument for P_{FA} (with P_{MD} following by a change of measure to the null). The upper bound is a similar to Thm 3.1. A similar argument shows $\Lambda_n(s)=1-s\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1\cup\mathcal{X}_\infty]\Theta(1)$, which provides an upper bound on the rate via the Chernoff bound. Note when $\mathsf{P}_{f_1}[\mathcal{X}_1]=o(\mathsf{P}_{f_1}[\mathcal{X}_\infty])$, Theorem 3.2 in [6] shows the $\Theta(1)$ quantity is at least 1.

For the lower bound, we proceed by changing to the same tilted measure used in Theorem 3.1. The main challenge in establishing the lower bound is estimating the contribution to the variance of the log-likelihood ratio of one sample under the tilted measure from $\mathcal{X}_\infty$. By convexity and differentiability of Λ_n (Lemma 2.2.5, [15]), $\Lambda_n'(s_n)=\mathsf{E}_0[\log(1+\epsilon_n(\mathsf{L}_n-1))(1+\epsilon_n(\mathsf{L}_n-1))^{s_n}]=0$. Using this relationship, we can approximate s_n. Using this approximation, we can show under the tilted measure, $\log(1+\epsilon_n(\mathsf{L}_n-1))$ has mean zero and variance $\sigma_n^2=\Theta(1)\epsilon_n^2D_n^2+\Theta(1)\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1]+\Theta(1)\log n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_\infty]$. Applying the assumption (9) allows the remainder of the proof to proceed similarly to Thm 3.1. ∎

Note that the conditions of the theorem are automatically satisfied so long as $\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1\cup\mathcal{X}_\infty]=\omega(\log n/n)$. As in the case of 3.1, we only require a logarithmic backoff from the detection limit given in Thm. 3.3.

Our final result provides conditions under which detection is impossible by analyzing the Hellinger distance as in [5].

Theorem 3.3: Consistent testing is possible if and only if $n\epsilon_n^2D_n^2\to\infty$ or $n\epsilon_n\mathsf{P}_{f_1}[\mathcal{X}_1\cup\mathcal{X}_\infty]\to\infty$. Moreover, if consistent testing is not possible, $\inf_{\delta_n}\mathsf{P}_{\mathrm{FA}}(n)+\mathsf{P}_{\mathrm{MD}}(n)\to 1$ where the infimum is taken over all collections of tests $\{\delta_n\}$.

Proof: Let f,g be two PMFs on $\mathcal{X}$, and define the Hellinger distance to be $H(f,g)=\sqrt{\sum_{\mathcal{X}}(\sqrt{f(x)}-\sqrt{g(x)})^2}$. The "if" direction is proved by the upper bounds in Thm. 3.1 and 3.2. The "only if" direction is proved by showing $H^2(f_{0,n},(1-\epsilon_n)f_{0,n}+\epsilon_nf_{1,n})=o(\frac{1}{n})$ if the consistency conditions are violated, which implies $\inf_{\delta_n}\mathsf{P}_{\mathrm{FA}}(n)+\mathsf{P}_{\mathrm{MD}}(n)\to 1$ as in [5]. ∎

The implications of this theorem are that whenever the LRT (5) is not consistent, no test gives better error probability than flipping a fair coin.

IV. Adaptive Testing

We consider the following test for when $f_{0,n}$ is known but $\{\epsilon_n,f_{1,n}\}$ are not:

$$\delta_n(X_1,\ldots,X_n)\triangleq\begin{cases}1 & D(\hat{p}_n||f_{0,n})\geq a_n\\ 0 & o.w.\end{cases} \quad (11)$$

where $\hat{p}_n(x)=\frac{\sum_{i=1}^n\mathbb{1}_{[X_i=x]}}{n}$ is the empirical distribution of the observations and $D(f||g)=\sum_{\mathcal{X}}f(x)\log\frac{f(x)}{g(x)}$ is the Kullback-Leibler (KL) divergence between f and g [14]. We will assume a_n is a sequence tending to zero such that $a_n>\frac{|\mathcal{X}|\log(n+1)}{n}$.

Note that the adaptive test only uses the empirical distribution, knowledge of the null distribution and the choice of a_n to make a decision (whereas the LRT (5) uses knowledge of ϵ_n and $f_{1,n}$ and therefore is not always practical). A larger choice of a_n improves the rate of false alarm at the cost of possibly excluding some possible $\{\epsilon_n,f_{1,n}\}$ pairs from being detected. Due to space constraints, we will assume $\frac{na_n}{\log n}\to\infty$. The adaptive test is a variant of Hoeffding's test [15].

Theorem 4.1: Assume $\mathcal{X}=\mathcal{X}_0$. Then, (11) satsifies:

1) (Known alternative hypothesis) $\lim_{n\to\infty}\frac{\log\mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2D_n^2}=-\frac{1}{8}$ if $a_n=\frac{\epsilon_n^2D_n^2}{8}(1+o(1))$
2) (Unknown alternative hypothesis) $\limsup_{n\to\infty}\frac{\log\mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2D_n^2}\leq-\frac{1}{2}$ if $a_n=o(\epsilon_n^2D_n^2)$

A. Proof Sketch for Rates for Adaptive Testing

Sanov's theorem [15] furnishes the following upper bound on the behavior of $\hat{p}_n$ lying in a set S when $X_1,\ldots,X_n$ are drawn i.i.d. from distribution g on finite alphabet $\mathcal{X}$:

$$\mathsf{P}_g[\hat{p}_n\in S]\leq(n+1)^{|\mathcal{X}|}e^{-n\inf_{v\in S}D(v||g)}. \quad (12)$$

There exist distributions where the behavior predicted by Sanov's theorem is essentially tight [17], so we conjecture that the performance of the test given by (11) cannot be refined significantly beyond what is presented in this work.

Applying Sanov's theorem shows that for the test (11),

$$\limsup_{n\to\infty}\frac{\log\mathsf{P}_{\mathrm{FA}}(n)}{na_n}\leq-1. \quad (13)$$

In order to study the rate under the alternative, we compute the exponent of (12) with $S=\{\delta_n=0\}$ and $g=(1-\epsilon_n)f_{0,n}+\epsilon_nf_{1,n}$, b_n. A standard argument (Problem 2.14, [18]) shows that the solution to

$$b_n=\inf_{v:D(v||f_{0,n})\leq a_n}D(v||(1-\epsilon_n)f_{0,n}+\epsilon_nf_{1,n}), \quad (14)$$

which specifies the rate under the alternative, is given by:

1) If $a_n\geq D((1-\epsilon_n)f_{0,n}+\epsilon_nf_{1,n}||f_{0,n})$, $b_n=0$.
2) If $a_n<D((1-\epsilon_n)f_{0,n}+\epsilon_nf_{1,n}||f_{0,n})$, then for some $\alpha_n\in(0,1]$,

$$(1-\alpha_n)\frac{\Lambda_n'(1-\alpha_n)}{\Lambda_n(1-\alpha_n)}-\log\Lambda_n(1-\alpha_n)=a_n \quad (15)$$

$$-\alpha_n\frac{\Lambda_n'(1-\alpha_n)}{\Lambda_n(1-\alpha_n)}-\log\Lambda_n(1-\alpha_n)=b_n. \quad (16)$$

By applying (12), we see (11) is consistent with rate given by

$$\limsup_{n\to\infty}\frac{\log\mathsf{P}_{\mathrm{MD}}(n)}{nb_n}\leq-1 \quad (17)$$

if $\frac{nb_n}{\log n}\to\infty$. The rate for the case where $\frac{nb_n}{\log n}=\Theta(1)$ is treated similarly to the case where $\frac{na_n}{\log n}\not\to\infty$.

V. Detection of Quantized Data

Many data acquisition systems operate by quantizing data to a fixed set of levels for storage or transmission. Prior to analyzing the data, the quantizer levels are reconstructed to approximate the original data and analysis designed for the original data is applied. This method of analysis can be suboptimal, since after quantization, all the information in the sample is encoded in the levels. Thus, testing based on the quantizer levels can lead to simpler (and possibly more powerful) tests than on the reconstructed data. An application is a sensor network quantizing a real-valued observation to send to a fusion center. We give some results based on our prior developed theory for detection of Gaussian mixtures via 1-bit quantization.

Consider the problem of detecting between

$$\mathsf{H}_{0,n} : X_1, \ldots, X_n \sim N(0,1) \text{ i.i.d.} \tag{18}$$

$$\mathsf{H}_{1,n} : X_1, \ldots, X_n \sim (1-\epsilon_n)N(0,1) + \epsilon_n N(\mu_n, 1) \text{ i.i.d.} \tag{19}$$

where $\epsilon_n = n^{-\beta}$ for some $\beta \in (0,1)$ and μ_n is a positive sequence. This mixture detection problem is known as the *Gaussian Location Model* (GLM).

The problem of when consistent tests exist has been well studied:

Theorem 5.1: ([2]–[4]) The boundary of the detectable region (in $\{(\epsilon_n, \mu_n)\}$ space) is given by (with $\epsilon_n = n^{-\beta}$):

1) If $0 < \beta \leq 1/2$, then $\mu_{crit,n} = n^{\beta - 1/2}$. (Dense)
2) If $1/2 < \beta < 3/4$, then $\mu_{crit,n} = \sqrt{2(\beta - \frac{1}{2}) \log n}$. (Moderately Sparse)
3) If $3/4 \leq \beta < 1$, then $\mu_{crit,n} = \sqrt{2(1-\sqrt{1-\beta})^2 \log n}$. (Very Sparse)

If $\mu_n > \mu_{crit,n}$ for all n sufficiently large, $\mathsf{P}_{\mathrm{FA}}(n) + \mathsf{P}_{\mathrm{MD}}(n) \to 0$ for the likelihood ratio test (5). Otherwise, any sequence of tests satisfies $\mathsf{P}_{\mathrm{FA}}(n) + \mathsf{P}_{\mathrm{MD}}(n) \to 1$.

In this section, we study the effects of quantization via the n-dependent quantizer

$$q_n(x) = \mathbb{1}_{\{x > c_n\}} \tag{20}$$

where c_n is a non-negative sequence.

Let $Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-x^2/2} dx$ denote the complementary normal cumulative distribution function. Then, the testing problem on quantized data is given by the following finite mixture detection problem:

$$\begin{aligned} \epsilon_n = n^{-\beta}, f_{0,n} &= [1 - Q(c_n), Q(c_n)] \\ f_{1,n} &= [1 - Q(c_n - \mu_n), Q(c_n - \mu_n)]. \end{aligned} \tag{21}$$

Our first result concerns fixed quantizers, where Thm 3.1 is applicable:

Theorem 5.2: Assume the quantizer is fixed independent of n, i.e. $c_n = c$. Let $D_n^2 = \frac{(Q(c) - Q(c-\mu_n))^2}{1-Q(c)} + \frac{(Q(c) - Q(c-\mu_n))^2}{Q(c)}$. Then, (5) applied to the quantized data is consistent if and only if $\beta < 1/2$ and $n\epsilon_n^2 D_n^2 \to \infty$. Moreover, the rate of (5) is given by

$$\lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n^2 D_n^2} = \lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{MD}}(n)}{n\epsilon_n^2 D_n^2} = -\frac{1}{8}. \tag{22}$$

The quantizer $c_n = 0$ leads to a consistent test for the entire dense detectable region, but the quantized has suboptimal rate compared to the unquantized test (since the quantizer does not differentiate between large and small x, when large x are more likely under the alternative) [6].

Our second result concerns quantizers whose levels can depend on n:

Theorem 5.3: Assume the quantizer is defined by the sequence $c_n = \sqrt{2 \log n}$. Then, for $\mu_n = \sqrt{2r \log n}$ where $(1 - \sqrt{1-\beta})^2 < r < 1$, the test specified by (5) applied to the quantized data is consistent and satisfies

$$\lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n Q(\sqrt{2\log n} - \mu_n)} = \lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{MD}}(n)}{n\epsilon_n Q(\sqrt{2\log n} - \mu_n)} = -1 \tag{23}$$

If $r > 1$ or $\mu_n = \omega(\sqrt{\log n})$, then

$$\lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{FA}}(n)}{n\epsilon_n} = \lim_{n\to\infty} \frac{\log \mathsf{P}_{\mathrm{MD}}(n)}{n\epsilon_n} = -1. \tag{24}$$

Otherwise, the test is not consistent.

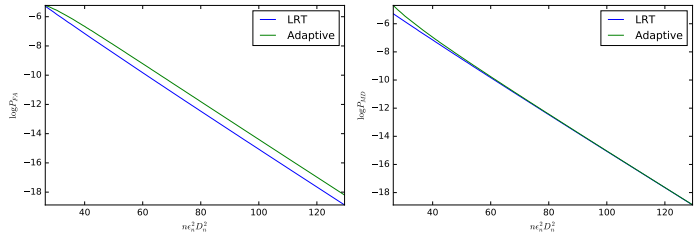
In this case, $\mathcal{X}_0 = \{0\}$ and $\mathcal{X}_\infty = \{1\}$ and Thm 3.2 can be applied. The threshold $c_n = \sqrt{2 \log n}$ corresponds to the mean of the maximum of a standard normal vector of length n. The detectable region is consistent with thresholding the sample maximum at level $\sqrt{2 \log n}$ for detection [3].

VI. Numerical Experiments

In this section, we illustrate our rate characterization by an example of 1-bit quantization of a Gaussian model. Consider the GLM given by (18) and (19) with $\epsilon_n = n^{-0.35}$ and $\mu_n = 2$. We study the performance of the the 1-bit quantizer with threshold $c_n = 0$ specified by (21). The rate characterization of the error probabilities for the LRT is stated in Thm 5.2. The rate characterization of the adaptive test proposed in Sec. IV is given in Thm 4.1.

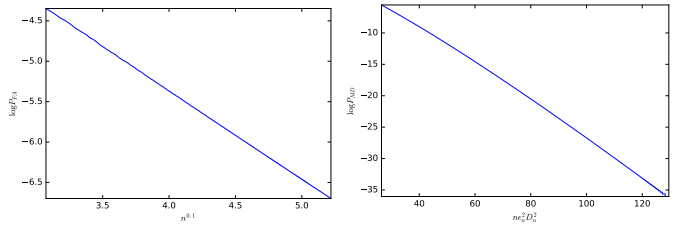
The performance of the LRT and adaptive test with threshold selected to match the false alarm rate of the LRT is shown in Fig. 1 for sample sizes up to 1.5×10^7. The error probabilities were computed exactly by noting that the error events are determined by the number of quantized samples that equal 1 (which follows a Binomial distribution under both hypotheses) and using the Binomial distribution function. We see that the slope of the log-error probabilities is -0.13 whereas the prediction of Thm. 5.2 and Thm. 4.1 is -0.125. Note that while the adaptive test has the same observed rate, its false alarm and miss detection probabilities are slightly higher than the LRT and the gap does not appear to grow with sample size. These results indicate that our theory accurate at reasonable sample sizes.

The performance of the adaptive test with an adaptive threshold selection $a_n = n^{-0.9}$ is given in Fig. 2 for sample sizes up to 1.5×10^7. The error probabilities were computed



(a) False Alarm (b) Miss Detection

Fig. 1: Simulations of error probabilities in the 1-bit quantized GLM (21) for $\epsilon_n = n^{-0.35}, \mu_n = 2$. The adaptive test (11) threshold is set to to match false alarm rate of the LRT (5).



(a) False Alarm (b) Miss Detection

Fig. 2: Simulations of error probabilities in the 1-bit quantized GLM (21) for $\epsilon_n = n^{-0.35}, \mu_n = 2$. The adaptive test (11) threshold is set to $a_n = n^{-0.9}$, independent of any knowledge of the alternative.

identically to the prior to example. We see the log-false alarm probability behaves as $-1.1n^{0.1}$ (which is close to the predicted $-n^{0.1}$), whereas the slope of the log-miss detection probability is $-0.31n\epsilon_n^2 D_n^2$ (versus the predicted $-0.5n\epsilon_n^2 D_n^2$). We expect better agreement with our theory for larger sample sizes. Note that while the false alarm probability is much higher in Fig. 2 than in the oracle threshold setting of Fig. 1, the adaptive threshold provides error probabilities that are small enough for most practical applications. The larger false alarm probabilities of the adaptive test allow for much smaller miss detection probabilities than the LRT in Fig. 1.

VII. Conclusions

In this work, we have presented an oracle rate analysis for error probabilities and adaptive test construction for detecting a sparse mixture of signal and noise from pure noise on a finite alphabet. Our adaptive test construction is competitive with the oracle test at reasonable sample sizes, and both tests have good agreement with our asymptotic predictions.

There are several interesting avenues of extension. One is the analysis of mixture detection problems on countable alphabets or growing finite alphabets. Some relevant large deviations results in this case are presented in [6], [17]. Another is analysis of other tests, such as a χ^2 goodness-of-fit test, which replaces the KL divergence in (11) with a χ^2-divergence. It is reasonable to expect based on Thm 4.1 that this test will have good rate performance as well in some cases. Based on Sec. V, we raise the question of how to design quantizers if detection is the primary goal, with only knowledge of the null distribution. This problem has been treated in related contexts [19]. Finally, restricting $\mathcal{F}$ to have some parametric structure may lead to some interesting extensions in the large-alphabet regime, as in [20].

References

[1] Y. Ingster and I. A. Suslina, *Nonparametric goodness-of-fit testing under Gaussian models*, vol. 169, Springer Science & Business Media, 2003.

[2] T. T. Cai, X. J. Jeng, and J. Jin, "Optimal detection of heterogeneous and heteroscedastic mixtures," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 73, no. 5, pp. 629–662, 2011.

[3] D. Donoho and J. Jin, "Higher criticism for detecting sparse heterogeneous mixtures," *Ann. Statist.*, vol. 32, no. 3, pp. 962–994, 06 2004.

[4] E. Arias-Castro and M. Wang, "Distribution-free tests for sparse heterogeneous mixtures," *arXiv preprint arXiv:1308.0346 [math.ST]*, 2013.

[5] T. T. Cai and Y. Wu, "Optimal detection of sparse mixtures against a given null distribution," *IEEE Trans. Info. Theory*, vol. 60, no. 4, pp. 2217–2232, 2014.

[6] J. G. Ligo, G. V. Moustakides, and V. V. Veeravalli, "Rate analysis for detection of sparse mixtures," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 4244–4248, Extended preprint arXiv:1509.07566 [cs.IT].

[7] 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 of Probability & Its Applications*, vol. 3, no. 2, pp. 161–173, 1958.

[8] M. R. Bloch, "Covert communication over noisy channels: a resolvability perspective," *IEEE Trans. Inform. Theory*, vol. 62, no. 5, pp. 2334–2354, 2016.

[9] J. Fridrich, *Steganography in digital media*, Cambridge University Press, Cambridge, 2010.

[10] E. Mossel and S. Roch, "Distance-based species tree estimation: information-theoretic trade-off between number of loci and sequence length under the coalescent," *arXiv preprint arXiv:1504.05289 [math.PR]*, 2015.

[11] J. J. Goeman and P. Bühlmann, "Analyzing gene expression data in terms of gene sets: methodological issues," *Bioinformatics*, vol. 23, no. 8, pp. 980–987, 2007.

[12] L. Cayon, J. Jin, and A. Treaster, "Higher criticism statistic: detecting and identifying non-Gaussianity in the WMAP first-year data," *Monthly Notices of the Royal Astronomical Society*, vol. 362, no. 3, pp. 826–832, 2005.

[13] D. Donoho and J. Jin, "Higher criticism thresholding: Optimal feature selection when useful features are rare and weak," *Proceedings of the National Academy of Sciences*, vol. 105, no. 39, pp. 14790–14795, 2008.

[14] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, NY: John Wiley and Sons, Inc., 2006.

[15] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, vol. 38, Springer Science & Business Media, 2009.

[16] F. den Hollander, *Large deviations*, vol. 14, American Mathematical Soc., 2008.

[17] W. C. M. Kallenberg, "On moderate and large deviations in multinomial distributions," *Ann. Statist.*, vol. 13, no. 4, pp. 1554–1580, 1985.

[18] I. Csiszár and J. Körner, *Information theory*, Cambridge University Press, Cambridge, second edition, 2011.

[19] H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey distance measures in the design generalized quantizers for binary decision systems," *IEEE Trans. Comm.*, vol. 25, no. 9, pp. 893–900, Sep 1977.

[20] J. Unnikrishnan et al., "Universal and composite hypothesis testing via mismatched divergence," *IEEE Trans. Inform. Theory*, vol. 57, no. 3, pp. 1587–1603, 2011.