RATE ANALYSIS FOR DETECTION OF SPARSE MIXTURES

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

ABSTRACT
In this paper, we study the rate of decay of the probability of error for distinguishing between a sparse signal with noise, modeled as a sparse mixture, from pure noise. This problem has many applications in signal processing, evolutionary biology, bioinformatics, astrophysics and feature selection for machine learning. We let the mixture probability tend to zero as the number of observations tends to infinity and derive oracle rates at which the error probability can be driven to zero for a general class of signal and noise distributions. In contrast to the problem of non-sparse signals, we see the log-probability of error decays sublinearly rather than linearly and is characterized through the \(\chi^2\)-divergence rather than the Kullback-Leibler divergence. This work provides the first characterization of the rate of decay of the error probability for this problem.

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

1. 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 many natural applications. In signal processing, it can be applied to detecting a signal in a multi-channel system or detecting covert communications [12]. In evolutionary biology, the problem manifests in the reconstruction of phylogenetic trees in the multi-species coalescent model [3]. In bioinformatics, detection of information in gene ontology datasets [4]. In astrophysics, detection of species, sparse detection, likelihood ratio test

Let \(\{f_{0,n}\}, \{f_{1,n}\}\) be sequences of probability density functions (PDFs) for real valued random-variables.

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

\[
H_{0,n} : X_1, \ldots, X_n \sim f_{0,n} \text{ i.i.d. (null)} \tag{1}
\]

\[
H_{1,n} : X_1, \ldots, X_n \sim (1 - \epsilon_n) f_{0,n} + \epsilon_n f_{1,n} \text{ i.i.d. (alternative)} \tag{2}
\]

where \(\{f_{0,n}\}\) is known and \(\{f_{1,n}\}\) is from some known family of sequences of PDFs \(\mathcal{F}\) and \(\{\epsilon_n\}\) is a sequence of positive numbers such that \(\epsilon_n \to 0\). We will also assume \(nc_n \to \infty\) so that a typical realization of the alternative is distinguishable from the null.

Let \(P_{0,n}, P_{1,n}\) denote the probability measure under \(H_{0,n}, H_{1,n}\) respectively, and let \(E_{0,n}, E_{1,n}\) be the corresponding expectations with respect to some known and fixed \(\{f_{0,n}\}, \{f_{1,n}\}\) and \(\{\epsilon_n\}\)

When convenient, we will drop the subscript \(n\). Let \(\Lambda_n \triangleq \frac{f_{1,n}}{f_{0,n}}\). When \(f_{0,n}(x) = f_0(x)\) and \(f_{1,n}(x) = f_0(x - \mu_n)\), we say that the model is a location model. When \(f_{0,n}\) is a standard normal PDF, we call the location model a Gaussian location model. The distributions of the alternative in a location model are described by the set of sequences \(\{\epsilon_n, \mu_n\}\).

The location model can be considered as one where the null corresponds to pure noise \(f_{0,n}\), while the alternative corresponds to a sparse signal (controlled by \(\epsilon_n\)) with signal strength \(\mu_n\) contaminated by additive noise. The relationship between \(\epsilon_n\) and \(\mu_n\) determines the signal-to-noise ratio (SNR), and characterizes when the hypotheses can be distinguished with vanishing probability of error. In the general case, \(f_{1,n}\) can be thought of as the signal distribution.

A hypothesis test \(\delta_n\) between \(H_{0,n}\) and \(H_{1,n}\) is a function \(\delta_n : \{x_1, \ldots, x_n\} \to \{0, 1\}\). We define the probability of false alarm for

\[\begin{align*}
\Pr_{H_{0,n}} (\delta_n = 1) &= c_n, \\
\Pr_{H_{1,n}} (\delta_n = 0) &= 1 - c_n
\end{align*}\]
a hypothesis test $\delta_n$ between $H_{0,n}$ and $H_{1,n}$ as

$$P_{FA}(n) \triangleq P_{0,n}(\delta_n = 1)$$

and the probability of missed detection as

$$P_{MD}(n) \triangleq P_{1,n}(\delta_n = 0).$$

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

$$\lim_{n \to \infty} \log P_{FA}(n) = -c \lim_{n \to \infty} \log P_{MD}(n) = -d$$

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 $\{7\}$. In the mixture detection problem, $g_0$ and $g_1$ will be sublinear functions of $n$.

The log-likelihood ratio between $H_{1,n}$ and $H_{0,n}$ is

$$LLR(n) = \sum_{i=1}^{n} \log (1 - \epsilon_n + \epsilon_n L_n(X_i)).$$

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

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

It is well known that $\{7\}$ is optimal in the sense of minimizing $P_{FA}(n) + P_{MD}(n)$ for testing between $H_{0,n}$ and $H_{1,n}$, which is the average probability of error when the null and alternative are assumed to be equally likely $\{15\}$. It is valuable to analyze $P_{FA}$ and $P_{MD}$ separately since many applications incur different penalties for false alarms and missed detections.

**Location Model:** The detectable region for a location model is the set of sequences $\{\epsilon_n, \mu_n\}$ such that a sequence of consistent oracle tests $\{\delta_n\}$ exist.

For convenience of analysis, we introduce the parameterization

$$\epsilon_n = n^{-\beta}$$

where $\beta \in (0, 1)$ as necessary. Following the terminology of $\{11\}$, when $\beta \in (0, 1/2)$ the mixture is said to be a “dense mixture”; if $\beta \in (1/2, 1)$, the mixture is said to be a “sparse mixture”.

### 2.1. Related Work

Prior work on mixture detection has been focused primarily on the Gaussian location model. The main goals in these works have been to determine the detectable region and construct optimally adaptive tests (i.e. those which are consistent independent of knowledge of $\{\epsilon_n, \mu_n\}$, whenever possible). The study of detection of mixtures where the mixture probability tends to zero was initiated by Ingster for the Gaussian location model $\{13\}$. Ingster characterized the detectable region, and showed that outside the detectable region the sum of the probabilities of false alarm and missed detection tends to one for any test. Since the generalized likelihood statistic tends to infinity under the null, Ingster developed an increasing sequence of simple hypothesis tests that are optimally adaptive. Donoho and Jin introduced the celebrated Higher Criticism test which is optimally adaptive and is computationally efficient, and also discussed some extensions to Subbotin distributions and $\chi^2$-distributions $\{2\}$. Cai et al. extended these results to the case where $f_{0,n}$ is standard normal and $f_{1,n}$ is a normal distribution with positive variance, derived limiting expressions for the distribution of $LLR(n)$ under both hypotheses, and showed that the Higher Criticism test is optimally adaptive in this case $\{9\}$. Jager and Wellner proposed a family of tests based on $\phi$-divergences and showed that they attain the full detectable region in the Gaussian location model $\{12\}$. Arias-Castro and Wang studied a location model where $f_{0,n}$ is some fixed but unknown symmetric distribution, and constructed an optimally adaptive test which relies only on the symmetry of the distribution when $\mu_n > 0$ $\{11\}$. Cai and Wu gave an information-theoretic characterization of the detectable region via an analysis of the sharp asymptotics of the Hellinger distance for a wide variety of distributions, and established a strong converse result showing that reliable detection is impossible outside the detectable region in many cases $\{10\}$. Our work complements $\{10\}$ by providing precise bounds on the error decay once the detectable region has been established. As shown in the next section, the error decay depends on the $\chi^2$-divergence between $f_{0,n}$ and $f_{1,n}$ rather than the Hellinger distance used in $\{10\}$. Since the $\chi^2$-divergence is not bounded above and below by the Hellinger distance in general $\{17\}$, our results cannot be derived from their methods.Walther numerically showed that while the popular Higher Criticism statistic is consistent, there exist optimally adaptive tests with significantly higher power for a given sample size at different sparsity levels $\{14\}$. Our work complements $\{14\}$ by providing a benchmark to meaningfully compare the sample size and sparsity tradeoffs of different tests with an oracle test. It should be noted that all work except $\{9,11\}$ has focused on the case where $\beta > 1/2$, and no prior work has provided an analysis of the rate at which $P_{FA}, P_{MD}$ can be driven to zero with sample size.

### 3. MAIN RESULTS FOR RATE ANALYSIS

#### 3.1. General Case

Our main result is a characterization of the oracle rate via the test given in $\{7\}$. The sufficient conditions required for the rate characterization are applicable to a broad range of parameters in the Gaussian location model $\{3.2\}$. Due to space constraints, we defer detailed proofs to $\{18\}$.

**Theorem 3.1 (\{13\})** Assume that for all $0 < \gamma < \gamma_0$ for some $\gamma_0 \in (0, 1)$, the following conditions are satisfied:

$$\lim_{n \to \infty} E_0 \left[ \frac{(L_n - 1)^2}{D_n^2} \mathbb{1}_{\{L_n \geq 1 + (\gamma/\epsilon_n)\}} \right] = 0$$

$$\epsilon_n D_n \to 0$$

$$\sqrt{n} \epsilon_n D_n \to \infty$$

where

$$D_n^2 = E_0[(L_n - 1)^2] < \infty.$$

Then for the test specified by $\{7\}$,

$$\lim_{n \to \infty} \frac{\log P_{FA}(n)}{n \epsilon_n^2 D_n^2} = \frac{1}{8}.$$

Moreover, $\{13\}$ holds replacing $P_{FA}$ with $P_{MD}$. 
The quantity $D_n^2$ is known as the $\chi^2$-divergence between $f_{0,n}$ and $f_{1,n}$. In contrast to the problem of testing between i.i.d. samples from two fixed distributions, the rate is not characterized by the Kullback-Leibler divergence for the mixture detection problem.

**Proof** (sketch) An upper bound on the rate for $P_{FA}$ is via the Chernoff bound with parameter $1/2$. The lower bound is established similarly to Cramer’s theorem (Thm I.4, [19]), which shows the Chernoff bound is tight for averages of i.i.d. random variables by a change of measure to a tilted distribution and the central limit theorem. We modify this proof by using a $n$-dependent tilting distribution and using the Lindeberg-Feller central limit theorem (Thm 2.4.5, [20]). The proof under the alternative is established by change of measure to the null.

When the conditions of Thm 3.1 do not hold, we have the following upper bound:

**Theorem 3.2** ([18]) If for all $M$ sufficiently large,

$$E_0 \left[ \log \left( \frac{f_{0,n}, \beta}{\log \frac{n}{n}} \right) \right] \rightarrow 1$$

then for the test specified by (7),

$$\limsup_{n \to \infty} \frac{\log P_{FA}(n)}{n} \leq -1.$$  \hspace{1cm} (15)

$$\lim_{n \to \infty} \frac{\log P_{MD}(n)}{n} = -1.$$  \hspace{1cm} (16)

Moreover, (15) holds with $P_{FA}$ replaced with $P_{MD}$.

**Proof** (sketch) The upper bounds on the rate are established via optimizing a Chernoff bound, where the optimal parameter tends to 0 or 1. The lower bound for the missed detection rate is by looking at the event where all mixture components under $H_{1,n}$ are drawn from $f_{0,n}$, and using consistency under $H_{0,n}$.

Interestingly, so long as the condition of Thm 3.2 holds, no non-trivial sequence of tests (i.e. $\limsup_{n \to \infty} P_{FA}, P_{MD} < 1$) can achieve a better rate than (7) under $H_{1,n}$. This is different from the case of testing i.i.d. observations from two fixed distributions, where allowing for large $P_{FA}$ can improve the rate under the alternative. It is reasonable to believe Thm 3.2 is tight for $P_{FA}$ as well, since $n\epsilon_n$ is the average number of signals under the alternative. However, we do not have a proof of this statement.

### 3.2. Gaussian Location Model

In this section, we apply Thms 3.1 and 3.2 to the Gaussian location model. The rate characterization proved is summarized in Fig. 1. Detailed proofs are given in [18]. We first recall some results from the literature for the detectable region for this model.

**Theorem 3.3** ([18],[17]) 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 $\mu_{\text{crit},n} = n^{\beta - 1/2}$. (Dense)
2. If $1/2 < \beta < 3/4$, then $\mu_{\text{crit},n} = \sqrt{2(1/2 - 1/2)} \log n$. (Moderately Sparse)
3. If $3/4 \leq \beta < 1$, then $\mu_{\text{crit},n} = \sqrt{2(1 - 1 - \beta) \log n}$. (Very Sparse)

If in the dense case $\mu_n = n^\beta$, then the LRT (7) is consistent if $r > \beta - 2$. Moreover, if $r < \beta - 1/2$, then $P_{FA}(n) + P_{MD}(n) \to 1$ for any sequence of tests as $n \to \infty$. If in the sparse cases, $\mu_n = \sqrt{2r \log n}$, then the LRT is consistent if $r > \mu_{\text{crit},n}$. Moreover, if $r < \beta - 2/2$, then $P_{FA}(n) + P_{MD}(n) \to 1$ for any sequence of tests as $n \to \infty$.

We call the set of $(\epsilon_n, \mu_n)$ sequences where (7) is consistent the interior of the detectable region. We now begin proving a rate characterization for the Gaussian location model by specializing Thm 3.3. Note that $L_n(x) = e^{\mu_n} \nu^{\beta/2} - D_n^2 = e^{\mu_n} - 1$.

**Corollary 3.4** (Dense case, [18]) If $\epsilon_n = n^{-\beta}$ for $\beta \in (0, 1/2)$ and $\mu_n = h(n)$ where $h(n) \to \infty$ and $\limsup_{n \to \infty} \frac{\mu_n}{\sqrt{2n \log n}} < 1$, then

$$\lim_{n \to \infty} \frac{\log P_{FA}(n)}{n} = -1.$$  \hspace{1cm} (17)

This result also holds replacing $P_{FA}$ with $P_{MD}$.

The implication of this corollary is that our rate characterization of the probabilities of error holds for a large portion detectable region up to the detection boundary, as $h(n)$ can be taken such that $h(n)/n^\beta \to 0$ for any $\xi > 0$, making it negligible with respect to $\mu_{\text{crit},n}$ in Thm 3.3.
Corollary 3.5 (Moderately sparse case) If $\epsilon_n = n^{-\beta}$ for $\beta \in (1/2, 3/4)$ and $\mu_n = \sqrt{2(\beta - 1/2 + \xi)} \log n$ for any $0 < \xi < (3/4) - 1/e\sqrt{2}$ then

$$\lim_{n \to \infty} \frac{\log P_{FA}(n)}{n\epsilon_n} = -\frac{1}{8}$$

and the same result holds replacing $P_{FA}$ with $P_{MD}$.

Note that $\xi$ can be replaced with an appropriately chosen sequence tending to 0.

For $\mu_n > \sqrt{2/\pi} \log n$, the first condition of Thm 3.1 does not hold. However, Thm 3.2 provides a partial rate characterization when $\mu_n > \sqrt{2\beta}\log n$:

Corollary 3.6 (Moderately sparse case) If $\epsilon_n = n^{-\beta}$ for $\beta \in (0, 1)$ and $\lim \inf_{n \to \infty} \frac{\log P_{FA}(n)}{n\epsilon_n} > 1$, then

$$\lim_{n \to \infty} \frac{\log P_{FA}(n)}{n\epsilon_n} < -1 \quad \text{and} \quad \lim_{n \to \infty} \frac{\log P_{MD}(n)}{n\epsilon_n} = -1. \quad (19)$$

Theorems 3.1 and 3.2 do not hold when $\epsilon_n = n^{-\beta}$ and $\mu_n = \sqrt{2r\log n}$ where $r \in (\beta/3, \beta)$ for $\beta \in (0, 3/4)$ or $r \in ((1 - \sqrt{1-\beta})^2, \beta)$ for $\beta \in (3/4, 1)$. An upper bound on the rate specific to the Gaussian location model is derived in [18] for this case.

4. NUMERICAL EXPERIMENTS

In this section, we provide numerical simulations to verify the rate characterization developed for the Gaussian location model. Due to space constraints, we focus on Cor. 3.4 and 3.5, as they correspond to the rates for the weakest signals which can be detected.

We first consider the dense case, with $\epsilon_n = n^{-0.4}$ and $\mu_n = 1$. Simulations were performed for $n = 10$ to $2 \times 10^7$ via direct monte carlo (10000 trials) or importance sampling (150000 trials) via the hypothesis alternate to the true hypothesis. The dashed lines are the best fit lines between the log-error probabilities and $n\epsilon_n^2(e^{\mu_n^2} - 1)$ using data for $n \geq 344000$. By Cor. 3.4, we expect the slope of the best fit lines to be approximately $-1/8$. As shown in Fig. 2a, the best-fit line corresponding to missed detection has slope $-0.1325$ with standard error (SE) 0.0038 ($-1/8$ is within 2 SE) and the line corresponding to false alarm has slope $-0.1240$ with standard error 0.00099 ($-1/8$ is within 1 SE). Thus, we have good agreement with Cor. 3.4.

The moderately sparse case with $\epsilon_n = n^{-0.6}$ and $\mu_n = \sqrt{2(0.19)\log n}$ is shown in Fig. 2b. Simulations were performed identically to the dense case. The dashed lines are the best fit lines between the log-error probabilities and $n\epsilon_n^2(e^{\mu_n^2} - 1)$ using data for $n \geq 100000$. By Cor. 3.5, we expect the slope of the best fit lines to be approximately $-0.125$. Both best fit lines have slope of $-0.108$ and standard error 0.002. It is important to note that $P_{FA}$, $P_{MD}$ are both large, even at $n = 2 \times 10^7$, and simulation to larger sample sizes should show better agreement with Cor. 3.5.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented an oracle rate characterization for the error probability decay with sample size in a general mixture detection problem. In the Gaussian location model, we explicitly showed that the rate characterization holds for a large portion of the dense regime and the moderately sparse regime. A partial rate characterization (an upper bound on the rate and universal lower bound on the rate under $H_{1n}$) was provided for the remainder of the detectable region. In contrast to usual large deviations results [7, 8] for the decay of error probabilities, our results show the log-probability of error decays sublinearly with sample size.

There are several possible extensions of this work. One is to provide corresponding lower bounds for the rate in cases not covered by Thm 3.1. Another is to provide a general analysis of the behavior that is not covered by Thm 3.1 and 3.2. As noted in [9], in some applications it is natural to require $P_{FA}(n) \leq \alpha$ for some fixed $\alpha > 0$, rather than requiring $P_{FA}(n) \to 0$. While Thm 3.3 shows the detectable region is not enlarged under in the Gaussian location model (and similarly for some general models [10]), it is conceivable that the oracle optimal test which fixes $P_{FA}$ (i.e. one which compares LLR(n) to a non-zero threshold) can achieve a better rate for $P_{MD}$. It is expected that the techniques developed in this paper extend to the case where $P_{FA}(n)$ is constrained to a level $\alpha$. Finally, it is important to develop tests that are amenable to a rate analysis and are computationally simple to implement over $0 < \beta < 1$. Some partial results for constructing such tests in the Gaussian location model via the Wasserstein distance are presented in [18]. Thresholding the largest observation is a candidate for rate analysis, though it is not consistent for a large portion of the detectable region [2].
6. REFERENCES


