

Classification with High-Dimensional Sparse Samples

Dayu Huang

Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign
Urbana, Illinois 61801
Email: dhuang8@illinois.edu

Sean Meyn

Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611
Email: meyn@ece.ufl.edu

Abstract—The task of the binary classification problem is to determine which of two distributions has generated a length- n test sequence. The two distributions are unknown; two training sequences of length N , one from each distribution, are observed. The distributions share an alphabet of size m , which is significantly larger than n and N . How does N, n, m affect the probability of classification error? We characterize the achievable error rate in a high-dimensional setting in which N, n, m all tend to infinity, under the assumption that probability of any symbol is $O(m^{-1})$. The results are:

- 1) There exists an asymptotically consistent classifier if and only if $m = o(\min\{N^2, Nn\})$. This extends the previous consistency result in [1] to the case $N \neq n$.
- 2) For the sparse sample case where $\max\{n, N\} = o(m)$, finer results are obtained: The best achievable probability of error decays as $-\log(P_e) = J \min\{N^2, Nn\}(1 + o(1))/m$ with $J > 0$.
- 3) A weighted coincidence-based classifier has non-zero generalized error exponent J .
- 4) The ℓ_2 -norm based classifier has $J = 0$.

Index Terms—high-dimensional model, large deviations, classification, sparse sample, generalized error exponent

I. INTRODUCTION

Consider the following binary classification problem: Two training sequences $\mathbf{X} = \{X_1, \dots, X_N\}$ and $\mathbf{Y} = \{Y_1, \dots, Y_N\}$ generated from two different *unknown* sources are observed. The two sources share the same alphabet $[m] := \{1, \dots, m\}$. Given a test sequence $\mathbf{Z} = \{Z_1, \dots, Z_n\}$, the classifier decides whether \mathbf{Z} comes from the first source or the second.

The performance of a classifier is usually assessed by how its probability of classification error depends on N, n, m . Since the exact formula for the probability of error is usually complicated, asymptotic models and performance criteria are used. For example, the classical error exponent criterion characterizes the exponential rate at which the probability of error decays as N and n increase to infinity. In addition to assessing a particular classifier's performance, it is desirable to establish fundamental limits on the best achievable performance.

The authors would like to acknowledge help discussions with Tuğkan Batu and Aaron Wagner.

Financial support from the National Science Foundation (NSF CCF 07-29031 and CCF 08-30776), ITMANET DARPA RK 2006-07284 and AFOSR grant FA9550-09-1-0190 is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NSF, DARPA or AFOSR.

In many applications such as text classification, the number of training and test samples observed, N and n , are much smaller than the size of alphabet m . This is the so-called *sparse sample* problem. For example, suppose we want to decide, given two articles written by two different authors, which author writes the third article. The number of words appearing in an article is much smaller than the English vocabulary, and the histogram of words is a sparse one [2].

The high-dimensional setting, in which N, n, m all tend to infinity and m is much larger than N, n , is a widely-used approach to analyze classifiers for the sparse sample problem. A widely-used performance criterion is asymptotic consistency: Given some dependence of N, n on m , does the probability of error decay to zero as m increases to infinity? A fundamental result with respect to this criterion was established in [1]: Assuming that the distribution on all symbols in the alphabet is of order $1/m$, there exists an asymptotic consistent classifier if and only if $m = o(n^2)$. Note that the result is established only for the case $N = n$.

In most practical scenarios, the number of test samples available is smaller than the number of training samples. It is thus desirable to understand how N and n affects the performance individually. We thus pose the following questions:

- 1) How fast do N and n need to increase with m in order to have an asymptotic consistent classifier?
- 2) Does the probability of error depend on N and n in the same way?
- 3) If the number of training samples is limited, can the performance be improved by having more test samples?

The goal of this paper is to answer these questions by establishing achievability and converse results on best achievable probability of classification error. Our tool is the generalized error exponent analysis technique from [3]. In this prior work, the sparse sample *goodness of fit* problem is investigated in which the number of test samples is much smaller than the size of alphabet. The classical error exponent was extended to this problem via a different scaling in large deviation analysis.

In the classification problem, the classical error exponent analysis has been applied to the case of fixed alphabet in [4] and [5]. It was shown that in order for the probability of error to decay exponentially fast with respect to n , the number of training samples N must grow at least linearly with n .

However, in the sparse sample problem, the classical error exponent concept is again not applicable, and thus a different scaling is needed.

We identify the appropriate scaling in this paper, and thereby obtain a generalized error exponent to approximate the probability of error for large but sparse observations. This analysis yields new insights on the best achievable performance:

- 1) The numbers of training and test samples N, n have different effects on the performance, made precise in Theorem IV.1 and Theorem IV.2.
- 2) The ℓ_2 -norm based classifier investigated in [1], which compares the ℓ_2 distances from the empirical distribution of the test sequence to those of the two training sequences, is sub-optimal in that it has zero generalized error exponent, while a weighted coincidence-based classifier proposed in this paper has a non-zero generalized error exponent.

Related work: Two problems that are closely related to the sparse sample classification problem is the goodness of fit problem and the problem of testing whether two distributions are close. For the goodness of fit problem, achievability and converse results with respect to different criteria have been established in [6], [7], [8], [9], [3]. For the problem of testing the closeness of two distributions, achievability and converse results with respect to asymptotic consistency have been established in [10], [11]. Our converse result uses the concept of profile in [12]. The results in [12] have lead to algorithms for classification and closeness testing [13], [14].

II. NOTATION AND MODEL

Consider the following classification problem: Two training sequences \mathbf{X} and \mathbf{Y} are generated i.i.d. with marginal distributions π and μ , respectively. Each symbol takes value in $[m] := \{1, 2, \dots, m\}$. A test sequence \mathbf{Z} is observed. The sequence \mathbf{Z} is i.i.d. with marginal distribution π under the null hypothesis $H0$ and with marginal μ under the alternative hypothesis $H1$. The three sequences $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ are independent.

Denote the set of probability distributions over $[m]$ by $\mathcal{P}([m])$. The pair of unknown distributions (π, μ) belongs to the following set $\Pi_m \subseteq \mathcal{P}([m]) \times \mathcal{P}([m])$,

$$\Pi_m = \{(\pi, \mu) : \|\mu - \pi\|_1 \geq \varepsilon, \max_j \pi_j \leq \frac{\eta}{m}, \max_j \mu_j \leq \frac{\eta}{m}\},$$

where η is a large positive constant. The definition of Π_m is essentially the same as the α -large-alphabet source defined in [1], except that we allow the number of training and test samples to be different. While this assumption that all words are rare does not hold for English vocabulary, the insights and classifiers obtained for rare words will be used to improve the algorithms for the case when there are both frequent and rare words.

The assumption that $\max_j \pi_j \leq \frac{\eta}{m}, \max_j \mu_j \leq \frac{\eta}{m}$ indicates that we are interested in how the existence of a large number of rare symbols affects the performance, and is motivated by the English vocabulary. Extending the results to the case where

there are both rare and non-rare symbols is a topic currently under investigation.

In the high-dimensional model, we consider a sequence of classification problems as described above, indexed by m . Thus $\mathcal{P}([m]), N, n, p, q, \Pi_m$ all depend on m . Moreover, N, n increase to infinity as m increases.

A classifier $\phi = \{\phi_m\}_{m \geq 1}$ is a sequence of binary-valued functions with $\phi_m : [m]^N \times [m]^N \times [m]^n \rightarrow \{0, 1\}$. It decides in favor of $H1$ if $\phi_m = 1$ and $H0$ otherwise. Use the notation $P_{(\mu, \pi, \nu)}(A)$ to denote the probability of the event A when \mathbf{X}, \mathbf{Y} and \mathbf{Z} have marginal distributions μ, π, ν respectively. The performance of a classifier ϕ is evaluated using the worst-case average probability of error given by

$$P_e(\phi_m) = \sup_{(\pi, \mu) \in \Pi_m} [\frac{1}{2} P_{(\pi, \mu, \pi)}\{\phi_m = 1\} + \frac{1}{2} P_{(\pi, \mu, \mu)}\{\phi_m = 0\}].$$

It is said to be asymptotically consistent if

$$\lim_{m \rightarrow \infty} P_e(\phi_m) = 0.$$

III. ASYMPTOTIC CONSISTENCY

We begin with the asymptotic consistency result.

Theorem III.1. *There exists an asymptotically consistent classifier if and only if*

$$m = o(\min\{N^2, Nn\}).$$

Proof: The sparse sample case where $\max\{N, n\} = o(m)$ is a corollary of the generalized error exponent analysis results given in Theorem IV.1 and Theorem IV.2.

Now consider the case when $m = O(N)$. The only if direction is trivial. For the if direction, when $m = o(N)$, the distributions of \mathbf{X} and \mathbf{Y} can be essentially be estimated with vanishing error since the number of types grows sub-exponentially in n (See [1, Lemma 3]). When m is linear in N , this problem can be transformed into a (harder) sparse sample problem with alphabet size mb where $b = \lceil \sqrt{\min\{N, n\}} \rceil$: Associate each symbol in $[m]$ with b symbols. Each observation is then randomly mapped to one of the associated symbols. A consistent classifier for the sparse sample problem leads to a consistent classifier for the original problem. ■

We have a few remarks:

- 1) For the case $N = n$, the conclusion of Theorem III.1 is consistent with the results in [1, Theorem 3 and 4]. Our proof technique is different.
- 2) The requirements on N and n for asymptotic consistency are different: The first requirement $m = o(N^2)$ needs to be satisfied regardless of how many test samples are available. The second requirement is active only when $n = O(N)$. Therefore, as long as the number of test samples grows linearly with the training samples, further increasing the test samples will not improve the performance in terms of asymptotic consistency.
- 3) On the other hand, increasing the number of training samples will always increase the performance. The effect of increasing the training samples is different when $n = o(N)$ and $N = o(n)$.

IV. GENERALIZED ERROR EXPONENT

When m is fixed, the following error exponent criterion has been used to evaluate a classifier ϕ :

$$I(\phi) := -\limsup_{n \rightarrow \infty} \frac{1}{n} \log(P_e(\phi_m)). \quad (1)$$

This classical error exponent criterion is no longer applicable in the sparse sample case where

Assumption 1. $N = o(m), n = o(m)$.

One should consider instead the following generalization, defined with respect to the normalization $r(N, n, m)$:

$$J(\phi) := -\limsup_{n \rightarrow \infty} \frac{1}{r(N, n, m)} \log(P_e(\phi_m)). \quad (2)$$

The results in Theorem IV.1 and Theorem IV.2 imply that the appropriate normalization is

$$r(N, n, m) = \min\{N^2, Nn\}/m.$$

The generalized error exponent $J(\phi)$ could depend on how N, n increase with m . Note that to have a consistent classifier, the necessary condition in Theorem III.1 must be satisfied, as summarized in the assumption below:

Assumption 2. $m = o(\min\{N^2, Nn\})$.

This is equivalent to $\lim_{m \rightarrow \infty} r(N, n, m) = \infty$.

The following theorems demonstrate that the definition in (2) is meaningful:

Theorem IV.1 (Achievability). *Suppose Assumption 1 and Assumption 2 hold. Then there exists a classifier ϕ such that*

$$J(\phi) > 0.$$

Theorem IV.2 (Converse). *Suppose Assumption 1 holds. There exists a constant \bar{J} such that for any classifier ϕ ,*

$$-\log(P_e(\phi_m)) \leq r(N, n, m)\bar{J}.$$

These theorems imply that the best achievable probability of error decays approximately as $P_e = \exp\{-r(N, n, m)J\}$ for some $J > 0$. Note that the probability of error changes exponentially with respect to n only when $n = O(N)$. When $N = o(n)$, the probability of error is mainly determined by the number of training samples. This phenomenon is similar to the case with fixed m , for which results in [4] show that whether $n = O(N)$ holds determines whether the probability of error decreases exponentially in n .

V. ℓ_2 -NORM BASED CLASSIFIER HAS A ZERO GENERALIZED ERROR EXPONENT

Let a_j^z be the number of times that j th symbol appears in \mathbf{Z} . The notations a^x and a^y are defined similarly.

The ℓ_2 -norm based classifier has the following test statistic:

$$F_n := \left\| \frac{1}{n} a^z - \frac{1}{N} a^x \right\|_2^2 - \left\| \frac{1}{n} a^z - \frac{1}{N} a^y \right\|_2^2.$$

The classifier is given by

$$\phi^F = \mathbb{I}\{F_n \geq 0\}.$$

This classifier was shown in [1] to be asymptotically consistent when $N = n$ and $m = o(N^2)$. We now show, however, this classifier has zero generalized error exponent:

Theorem V.1. *Suppose Assumption 1 and Assumption 2 hold and $N = n$. Assume in addition that $m = o(n^2/\log(n)^2)$. Then*

$$J(\phi^F) = 0.$$

The sub-optimality of ϕ^F is due to the following reason: For any j , a large variation of the value of a_j^y causes a significant change in the value of the statistic F_n . Assume m is even for simplicity of exposition. Let u denote the uniform distribution on $[m]$. Let $q_j = (1 + \varepsilon)/m$ for $j \leq m/2$ and $q_j = (1 - \varepsilon)/m$ for $j > m/2$. Consider the case where under H_0 , the distribution is given by (q, u, q) .

Considering the following event where one symbol appears many times:

$$C_n := \{a_1^y = \lfloor 4n/\sqrt{m} \rfloor\}, \quad (3)$$

we claim that this event is likely to cause a false alarm:

$$P_{(q,u,q)}\{\phi^F = 1 | C_n\} = 1 - o(1).$$

On the other hand, the probability of C_n decays slowly:

$$P_{(q,u,q)}(C_n) = \exp\{-4(n/\sqrt{m})\log(m)(1 + o(1))\}. \quad (4)$$

Combining these two equality gives the lower-bound

$$\begin{aligned} \log(P_e(\phi^F)) &\geq \log\left(\frac{1}{2}P_{(q,u,q)}(C_n)P_{(q,u,q)}\{\phi^F = 1 | C_n\}\right) \\ &= 34\frac{n}{\sqrt{m}}\log(m)(1 + o(1)) \end{aligned}$$

Thus this error decays at most as $nm^{-\frac{1}{2}}\log(m)$, slower than n^2/m . Consequently, $J(\phi^F) = 0$.

VI. PROOF OF ACHIEVABILITY: WEIGHTED COINCIDENCE-BASED CLASSIFIER

A nonzero generalized error exponent is achieved by the following weighted coincidence-based classifier, whose construction is inspired by the weighted coincidence-based test proposed in [3]. Define the test statistic T_n :

$$\begin{aligned} T_n = \sum_j &\left[\frac{1}{N^2} \mathbb{I}\{a_j^x = 2, a_j^z = 0\} + \frac{1}{n^2} \mathbb{I}\{a_j^x = 0, a_j^z = 2\} \right. \\ &- \frac{1}{nN} \mathbb{I}\{a_j^x = 1, a_j^z = 1\} + \frac{1}{nN} \mathbb{I}\{a_j^y = 1, a_j^z = 1\} \\ &\left. - \frac{1}{n^2} \mathbb{I}\{a_j^y = 0, a_j^z = 2\} - \frac{1}{N^2} \mathbb{I}\{a_j^y = 2, a_j^z = 0\} \right]. \end{aligned}$$

The classifier is given by $\phi^T = \mathbb{I}\{T_n \geq 0\}$.

Theorem IV.1 is proved by bounding $P_e(\phi^T)$ via Chernoff:

$$\log(P_{(\pi,\mu,\pi)}\{\phi^T = 1\}) \leq \inf_{\theta} \Lambda_{(\pi,\mu,\pi)}(\theta).$$

$$\log(P_{(\pi,\mu,\mu)}\{\phi^T = 0\}) \leq \inf_{\theta} \Lambda_{(\pi,\mu,\mu)}(\theta).$$

where $\Lambda_{(\pi,\mu,\nu)}(\theta) = \log E_{(\pi,\mu,\nu)}[\exp(\theta K_n)]$ is the logarithmic moment generating function of K_n . The main step is to obtain an asymptotic approximation to $\Lambda_{(\pi,\mu,\nu)}(\theta)$, given in the following proposition:

Proposition VI.1. Let $\theta = \min\{N^2, nN\}\gamma$. For $\gamma = O(1)$,

$$\begin{aligned} & \Lambda_{(\pi, \mu, \nu)}(\theta) \\ & \leq \min\{N^2, nN\} \left(\gamma \left[\sum_{j=1}^m \left(\frac{1}{2}(\pi_j - \nu_j)^2 - \frac{1}{2}(\mu_j - \nu_j)^2 \right) \right] \right. \\ & \quad \left. + \gamma^2 \left[\sum_{j=1}^m (\pi_j \nu_j + \mu_j \nu_j) + \frac{1}{2}(\pi_j^2 + \mu_j^2) \right] \right) \\ & \quad + O\left(\frac{\min\{N^2, nN\} \max\{N, n\}}{m^2}\right) + O(1). \end{aligned}$$

Proposition VI.1 is obtained using the Poissonization technique: The distribution of the vector a_j^x is the same as the conditional distribution of a vector of Poisson random variables whose expected values are given by $\lambda\pi$ for some constant $\lambda > 0$, conditioned on the event that the sum of these random variables is equal to N . The main steps are similar to those used for results in [3].

Applying Proposition VI.1 with the Chernoff bound for the cases $\nu = \pi$ and $\nu = \mu$, and using Assumption 1 and Assumption 2, and the facts $\pi_j, \mu_j \leq \eta/m$ and $\sum_{j=1}^m (\mu_j - \pi_j)^2 \geq \varepsilon^2/m$, we obtain

$$\begin{aligned} \log(\mathbb{P}_{\pi, \mu, \pi}\{\phi^T = 1\}) & \leq -\frac{\varepsilon^4}{160\eta^2} \frac{\min\{N^2, nN\}}{m} (1 + o(1)), \\ \log(\mathbb{P}_{\pi, \mu, \mu}\{\phi^T = 0\}) & \leq -\frac{\varepsilon^4}{160\eta^2} \frac{\min\{N^2, nN\}}{m} (1 + o(1)). \end{aligned}$$

Note that the approximation $o(1)$ is uniform over all $(\pi, \mu) \in \Pi_m$. Therefore,

$$J \geq \frac{\varepsilon^4}{160\eta^2}.$$

VII. PROOF OF CONVERSE

Step 1: Establish the upper bound,

$$-\log(P_e(\phi_m)) \leq \bar{J}_1 N^2/m. \quad (5)$$

The main idea of the proof is to consider a event under which observations do not give any information regarding the hypotheses, and lower-bound the probability of such a event.

We now make this precise. Define the event

$$A = \{\text{No symbol in } \mathbf{X} \text{ appears more than once;} \\ \text{no symbol in } \mathbf{Y} \text{ appears more than once.}\}$$

Assume without loss of generality that m is even. Define a collection of bi-uniform distributions as follows: Let K_m denote the collection of all subsets of $[m]$ whose cardinality is $m/2$. For each set $\omega \in K_m$, define the distribution q^ω as

$$q_j^\omega = \begin{cases} (1 + \varepsilon)/m, & j \in \omega; \\ (1 - \varepsilon)/m, & j \in [m] \setminus \omega. \end{cases} \quad (6)$$

Note that $\|u - q^\omega\|_1 = \varepsilon$, and $(u, q^\omega) \in \Pi_m$ for all ω .

We will use the short-hand notation $\{(x, y, z)\} = \{(\mathbf{X}, \mathbf{Y}, \mathbf{Z}) = (x, y, z)\}$ throughout the paper.

Our choice of the collection of distributions makes sure that the following result holds:

Lemma VII.1. For any sequence $(x, y, z) \subseteq A$,

$$\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(x, y, z) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(q^\omega, u, u)}(x, y, z).$$

Proof sketch for Lemma VII.1: For any sequence, let φ_i denote the number of symbols appearing i times. The vector $[\varphi_1, \varphi_2, \varphi_3, \dots]$ is called the *profile* of the sequence [12].

Because of the symmetry of the collection of distributions $\{q^\omega, \omega \in K_m\}$, the symmetry of the uniform distribution u , and the independence among $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$, the value of $\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(x, y, z)$ only depends on the profiles of x, y , and z . In the event A , the profiles of x and y are fixed, which then leads to the claim of the lemma. ■

Lemma VII.1 implies that for any observation $(x, y, z) \in A$, it is impossible to tell whether it is more likely to come from the mixture on the left-hand side or the mixture on the right-hand side. Consequently,

$$\begin{aligned} & P_e(\phi_m) \\ & \geq \frac{1}{4|K_m|} \sum_{\omega} [\mathbb{P}_{(u, q^\omega, u)}\{\phi_m = 1\} + \mathbb{P}_{(u, q^\omega, q^\omega)}\{\phi_m = 0\}] \\ & \quad + \frac{1}{4|K_m|} \sum_{\omega} [\mathbb{P}_{(q^\omega, u, q^\omega)}\{\phi_m = 1\} + \mathbb{P}_{(q^\omega, u, u)}\{\phi_m = 0\}] \\ & \geq \frac{1}{4|K_m|} \sum_{\omega} \left[\Pr_{(u, q^\omega, u)}\{\phi_m = 1\} + \Pr_{(q^\omega, u, u)}\{\phi_m = 0\} \right] \\ & \geq \frac{1}{4|K_m|} \sum_{\omega} \left[\Pr_{(u, q^\omega, u)}(\{\phi_m = 1\} \cap A) + \Pr_{(q^\omega, u, u)}(\{\phi_m = 0\} \cap A) \right] \\ & = \frac{1}{4|K_m|} \sum_{\omega} \left[\Pr_{(u, q^\omega, u)}(\{\phi_m = 1\} \cap A) + \Pr_{(u, q^\omega, u)}(\{\phi_m = 0\} \cap A) \right] \\ & = \frac{1}{4|K_m|} \sum_{\omega} \Pr_{(u, q^\omega, u)}(A). \end{aligned} \quad (7)$$

where the first inequality follows from the fact that the maximum is no smaller than the average, and the second last inequality follows from Lemma VII.1. The probability of the event A can be lower-bounded.

Lemma VII.2. The following approximations holds uniformly for any ω :

$$\log\left(\Pr_{(u, q^\omega, u)}(A)\right) = -(1 + \frac{1}{2}\varepsilon^2) \frac{N^2}{m} (1 + o(1)) + O(1).$$

Proof sketch: It follows from a combinatorial argument that the probability that no symbol appears twice in \mathbf{X} when \mathbf{X} has marginal distribution u is given by

$$m(m-1) \dots (m-N+1)(1/m)^N = \exp\left\{-\frac{1}{2} \frac{N^2}{m} (1 + o(1))\right\}.$$

Estimating the probability that no symbol appears twice in \mathbf{Y} can be done similarly but is more involved. ■

The claim (5) follows from applying Lemma VII.2 to (7), and picking a large enough \bar{J} .

Step 2: Establish the second upper-bound

$$-\log(P_e(\phi_m)) \leq \bar{J}_2 (Nn + n^2)/m. \quad (8)$$

We consider the following event:

$$B = \{\text{No symbol in } \mathbf{Z} \text{ appears more than once;} \\ \text{no symbol in } \mathbf{Z} \text{ has appeared in either } \mathbf{X} \text{ or } \mathbf{Y}\}.$$

When this event happens, it is impossible (in the worst-case setting) to infer which distribution the test sequence is more likely to be generated from. This is captured by the following lemma:

Lemma VII.3. *Consider any \mathbf{x}, \mathbf{y} . For any two sequences \mathbf{z} and $\bar{\mathbf{z}}$ such that $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \subseteq B$ and $(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}) \subseteq B$, the following holds:*

$$\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, u)}(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}). \\ \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, q^\omega)}(\mathbf{x}, \mathbf{y}, \mathbf{z}) = \frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, q^\omega)}(\mathbf{x}, \mathbf{y}, \bar{\mathbf{z}}).$$

Proof sketch for Lemma VII.3: Since no symbols in \mathbf{z} have appeared in \mathbf{x} and \mathbf{y} , due to the symmetry of the collection of distributions $\{q^\omega, \omega \in K_m\}$ and the symmetry of the uniform distribution u , for fixed \mathbf{x} and \mathbf{y} , the value of $\frac{1}{|K_m|} \sum_{\omega \in K_m} \Pr_{(u, q^\omega, q^\omega)}(\mathbf{x}, \mathbf{y}, \mathbf{z})$ only depends on the profile of \mathbf{z} . It follows from the definition of the event B that the profile of \mathbf{z} is the same as the profile of $\bar{\mathbf{z}}$. ■

The result of Lemma VII.3 can be interpreted as follows: In the event B , observing \mathbf{Z} does not give any information since under either hypothesis, each sequence \mathbf{z} appears with equal probability.

Consider any \mathbf{x}, \mathbf{y} . Let $D\mathbf{x}, \mathbf{y} = \{\mathbf{z} : (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \{\phi_m = 1\} \cap B\}$ and $D^c\mathbf{x}, \mathbf{y} = \{\mathbf{z} : (\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \{\phi_m = 0\} \cap B\}$. Lemma VII.3 implies that the probability of $\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \phi_m = 1\} \cap B$ only depends on the size of $D\mathbf{x}, \mathbf{y}$, rather than what sequences the set $D\mathbf{x}, \mathbf{y}$ includes. Consequently, We then have

$$\begin{aligned} & \frac{1}{|K_m|} \sum_{\omega} \left[\Pr_{(u, q^\omega, u)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \phi_m = 1\} \cap B) \right. \\ & \quad \left. + \Pr_{(u, q^\omega, q^\omega)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}, \phi_m = 0\} \cap B) \right] \\ &= \left[\frac{1}{|K_m|} \sum_{\omega} \Pr_{(u, q^\omega, u)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B) \right] \frac{|D\mathbf{x}, \mathbf{y}|}{|D\mathbf{x}, \mathbf{y}| + |D^c\mathbf{x}, \mathbf{y}|} \\ & \quad + \left[\frac{1}{|K_m|} \sum_{\omega} \Pr_{(u, q^\omega, q^\omega)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B) \right] \frac{|D^c\mathbf{x}, \mathbf{y}|}{|D\mathbf{x}, \mathbf{y}| + |D^c\mathbf{x}, \mathbf{y}|} \\ &\geq \frac{1}{|K_m|} \min \left\{ \sum_{\omega} \Pr_{(u, q^\omega, u)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B), \right. \\ & \quad \left. \sum_{\omega} \Pr_{(u, q^\omega, q^\omega)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B) \right\}, \end{aligned} \tag{9}$$

where the inequality follows from lower-bounding the probability of $\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B$ under (u, q^ω, u) and (u, q^ω, q^ω) by the minimum of these two.

Lemma VII.4. *Let $\bar{J}_2 = 5$. Then the following bounds hold uniformly over all $\omega, \mathbf{x}, \mathbf{y}$:*

$$\log \left[\frac{\Pr_{(u, q^\omega, u)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B)}{\Pr_{(u, q^\omega, u)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\})} \right] \geq \bar{J}_2 \frac{Nn + n^2}{m} (1 + o(1)).$$

$$\log \left[\frac{\Pr_{(u, q^\omega, q^\omega)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\} \cap B)}{\Pr_{(u, q^\omega, q^\omega)}(\{\mathbf{X} = \mathbf{x}, \mathbf{Y} = \mathbf{y}\})} \right] \geq \bar{J}_2 \frac{Nn + n^2}{m} (1 + o(1)).$$

The proof is similar to that of Lemma VII.2.

Note that the average probability of error is equal to the summation of the left-hand side of (9) over all possible (\mathbf{x}, \mathbf{y}) . Applying Lemma VII.4 to lower-bound the right-hand side of (9) leads to the claim.

We now combine (5) and (8). It is straightforward to verify that

$$\min\{N^2, Nn + n^2\} \leq \min\{N^2, 2Nn\}.$$

Taking $\bar{J} = \max\{\bar{J}_1, 2\bar{J}_2\}$ leads to the claim of the theorem.

VIII. CONCLUSIONS AND FUTURE WORK

We have investigated the binary classification problem with sparse samples using generalized error exponent concept, and established fundamental performance limits. We have proposed a classifier that performs better than the ℓ_2 -norm based classifier. Future directions include:

- 1) Investigate classification algorithms that are applicable when there are both rare and frequent symbols.
- 2) The generalized error exponent analysis could be applicable to the problem of testing closeness of distributions.

REFERENCES

- [1] B. G. Kelly, T. Tularak, A. B. Wagner, and P. Viswanath, "Universal hypothesis testing in the learning-limited regime," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, Jun. 2010, pp. 1478 – 1482.
- [2] T. Zhang and F. Oles, "Text categorization based on regularized linear classification methods," *Information Retrieval*, vol. 4, pp. 5 – 31, 2001.
- [3] D. Huang and S. Meyn, "Error exponents for composite hypothesis testing with small samples," 2012, accepted for presentation at 2012 International Conference on Acoustic, Speech and Signal Processing.
- [4] J. Ziv, "On classification with empirically observed statistics and universal data compression," *IEEE Transactions on Information Theory*, vol. 34, no. 2, pp. 278 – 286, Mar. 1988.
- [5] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401 – 408, Mar. 1989.
- [6] T. Batu, E. Fischer, L. Fortnow, R. Kumar, R. Rubinfeld, and P. White, "Testing random variables for independence and identity," in *Proceedings of 42nd IEEE Symposium on Foundations of Computer Science*, Oct. 2001, pp. 442 – 451.
- [7] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Transactions on Information Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct. 2008.
- [8] M. S. Ermakov, "Asymptotic minimaxity of chi-square tests," *Theory of Probability and its Applications*, vol. 42, p. 589, 1998.
- [9] A. R. Barron, "Uniformly powerful goodness of fit tests," *The Annals of Statistics*, vol. 17, no. 1, pp. 107 – 124, 1989.
- [10] T. Batu, L. Fortnow, R. Rubinfeld, W. D. Smith, and P. White, "Testing that distributions are close," in *Proceedings of 41st Annual Symposium on Foundations of Computer Science*, 2000, pp. 259 – 269.
- [11] P. Valiant, "Testing symmetric properties of distributions," in *Proceedings of the 40th Annual ACM symposium on Theory of Computing*. New York, NY, USA: ACM, 2008, pp. 383 – 392.
- [12] A. Orlitsky, N. P. Santhanam, and J. Zhang, "Universal compression of memoryless sources over unknown alphabets," *IEEE Transactions on Information Theory*, vol. 50, no. 7, pp. 1469 – 1481, Jul. 2004.
- [13] J. Acharya, H. Das, A. Orlitsky, S. Pan, and N. P. Santhanam, "Classification using pattern probability estimators," in *Proceedings of 2010 IEEE International Symposium on Information Theory*, Austin, TX, Jun. 2010, pp. 1493 – 1497.
- [14] J. Acharya, H. D. A. Jafarpour, A. Orlitsky, and S. Pan, "Competitive closeness testing," in *Proceedings of 24th Annual Conference on Learning Theory*, Budapest, Hungary, Jun. 2011, pp. 47–68.