OPTIMALITY OF COINCIDENCE-BASED GOODNESS OF FIT TEST FOR SPARSE SAMPLE PROBLEMS

Dayu Huang and Sean Meyn

CSL & ECE
University of Illinois at Urbana-Champaign
1308 West Main Street, Urbana, IL 61801, USA

ABSTRACT

We consider the sparse sample goodness of fit problem, where the number of samples n is smaller than the size of the alphabet m. The generalized error exponent based on large deviation analysis was proposed to characterize the performance of tests, using the high-dimensional model in which both n and m tend to infinity and $n = o(m)$. In previous work, the best achievable probability of error is shown to decay $-\log(P_e) = (n^2/m)(1 + o(1))J$ with upper and lower bounds on J. However, there is a significant gap between the two bounds.

In this paper, we close the gap by proving a tight upper-bound, which matches the lower-bound over the entire region of generalized error exponents of false alarm and missed detection, achieved by the coincidence-based test. This implies that the coincidence-based test is asymptotically optimal.

Index Terms— chi-square test, high-dimensional model, goodness of fit, large deviations, optimal test

1. INTRODUCTION

Goodness of fit problem with small number of samples arises from many applications such as biomedical research and social science where the cost of obtaining a sample is high. To evaluate a test for the case when the number of samples n is smaller than the size of alphabet m, the criterion of *generalized error exponent* was proposed in the previous work [1]: it characterizes the rate that the probability of error converges to zero in a high-dimensional model where n and m both increase to infinity and $n = o(m)$. This criterion provides insights that are not available from asymptotic consistency analysis or Central Limit Theorem analysis: The widely used Pearson's chi-square test has a zero generalized error exponent of probability of error while a coincidence-based test proposed in [2] has a non-zero generalized error exponent.

Lower and upper bounds on the best achievable generalized error exponent have been obtained in previous works but they are not tight. It remained an open question what the optimal test is.

In this paper, answer this question by showing that there exists a better upper-bound that matches the existing lower-bound achieved by the coincidence-based test, and thus the coincidence-based test is optimal.

The technique used in proof of previous upper-bound only allows us to bound the probability of *missed detection* while leaving out the probability of *false alarm*. The proof of the new result uses a technique that allows us to *simultaneously* bound the generalized error exponents of false alarm and missed detection. We believe this technique could find its application in statistical inference problems where tight hardness results on rate of convergence are desirable.

1.1. Problem statement

Consider the following goodness of fit problem: Suppose the observations take value in the finite alphabet $[m] := \{1, 2, \ldots, m\}$, and denote the set of probability distribution over $[m]$ by $\mathcal{P}([m])$. An i.i.d. sequence $\boldsymbol{Z}_1^n = \{Z_1, \ldots, Z_n\}$ is observed, where $Z_i \in [m]$. Under the null hypothesis $H0$, Z_i has a *uniform* distribution π over $[m]$; under the alternative hypothesis $H1$, Z_i has a unknown distribution $\mu \in \Pi_m$, where Π_m is given by

$$\Pi_m = \{\mu : d(\mu, \pi) \geq \varepsilon\}, \tag{1}$$

and d is the total-variation metric:

$$d(\mu, \pi) = \sup\{|\mu(A) - \pi(A)| : A \subseteq [m]\} = \tfrac{1}{2}\|\mu - \pi\|_1.$$

A test $\phi = \{\phi_n\}_{n\geq 1}$ is a sequence of binary-valued function $\phi_n : [m]^n \to \{0, 1\}$. It decides in favor of $H1$ if $\phi_n = 1$ and $H0$ otherwise. The performance of a test is evaluated using the probability of false-alarm P_F and (worst-case) probability of missed detection P_M:

$$P_F(\phi_n) = \mathsf{P}_\pi\{\phi_n = 1\}, \quad P_M(\phi_n, \mu) = \mathsf{P}_\mu\{\phi_n = 0\},$$

$$P_M(\phi_n) = \sup_{\mu \in \Pi_m} P_M(\phi_n, \mu).$$

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.

The following generalization of classical error exponents has been proposed, defined with respect to the normalization $r(n,m)$ which takes value $r(n,m) = n^2/m$ when $n = o(m)$:

$$\begin{aligned} J_F(\phi, \mathbf{m}) &:= -\limsup_{n\to\infty} r^{-1}(n,m)\log(P_F(\phi_n)), \\ J_M(\phi, \mathbf{m}) &:= -\limsup_{n\to\infty} r^{-1}(n,m)\log(P_M(\phi_n)). \end{aligned} \tag{2}$$

Note that in the classical error exponent applicable to the case $m = O(n)$ is defined with the normalization is $r(n,m) = n$.

1.2. Previous results

The following lower and upper bounds on (J_F, J_M) have been established in [1]:

Theorem 1.1 (Lower-bound from achievability result in [1])**.** *The following pair of generalized error exponents are achieved by the coincidence-based test ϕ^{K}: For $\tau \in [0, \kappa(\varepsilon)]$,*

$$J_F(\phi^{\mathsf{K}}) = \sup_{\theta\geq 0}\{\theta\tau - \tfrac{1}{2}(e^{2\theta} - 1 - 2\theta)\},$$
$$J_M(\phi^{\mathsf{K}}) = \sup_{\theta\geq 0}\{\theta(\kappa(\varepsilon) - \tau) - \tfrac{1}{2}(e^{-2\theta} - 1 + 2\theta)(1 + \kappa(\varepsilon))\}$$

where

$$\kappa(\varepsilon) = \begin{cases} \frac{\varepsilon}{1-\varepsilon}, & \varepsilon \geq 0.5, \\ 4\varepsilon^2, & \varepsilon < 0.5. \end{cases} \tag{3}$$

The coincidence-based test that achieves this pair of generalized error exponents was introduced in [2]: Let

$$K_n = \sum_{j=1}^{m} \mathbb{I}\{n\Gamma_j^n = 1\}, \tag{4}$$

where $\Gamma_j^n = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}\{Z_i = j\}$ is the empirical distribution. This test statistic counts the number of symbols appearing once. The test is given by $\phi^{\mathsf{K}} = \mathbb{I}\{K_n \leq \mathsf{E}_{\pi^n}[K_n] - \tau_n\}$.

Theorem 1.2 (Upper-bound from hardness result in [1])**.** *For any test ϕ satisfying*

$$\lim_{n\to\infty} P_F(\phi_n) = 0, \tag{5}$$

the following upper-bound on the generalized error exponent of missed detection holds:

$$J_M(\phi, \mathbf{m}) \leq \bar{J}(\varepsilon). \tag{6}$$

where

$$\bar{J}(\varepsilon) = \sup_{\theta\geq 0}\{\theta\kappa(\varepsilon) - \tfrac{1}{2}(e^{-2\theta} - 1 + 2\theta)(1 + \kappa(\varepsilon))\} \tag{7}$$

The right-hand side of (7) is equal to the value of $J_M(\phi^K)$ given in Theorem 1.1 with $\tau = 0$. Therefore, the upper-bound on J_M given in (6), which holds for any J_F, is only tight at $J_F = 0$. It is desirable to obtain an upper-bound on the generalized error exponent that is tight over the entire region.

2. MAIN RESULTS

Theorem 2.1 (Tight upper-bound)**.** *Consider any $\tau \in [0, \kappa(\varepsilon)]$. For any test ϕ satisfying*

$$J_F(\phi) \geq \sup_{\theta\geq 0}\{\theta\tau - \tfrac{1}{2}(e^{2\theta} - 1 - 2\theta)\}, \tag{8}$$

the following upper-bound on the generalized error exponent of missed detection hold:

$$J_M(\phi) \leq \sup_{\theta\geq 0}\{\theta(\kappa(\varepsilon) - \tau) - \tfrac{1}{2}(e^{-2\theta} - 1 + 2\theta)(1 + \kappa(\varepsilon))\}. \tag{9}$$

The achievability and new hardness results are shown in Fig. 1. The upper-bound matches the lower-bound over the entire region , and thus the coincidence-based test is optimal with respect to generalized error exponent.

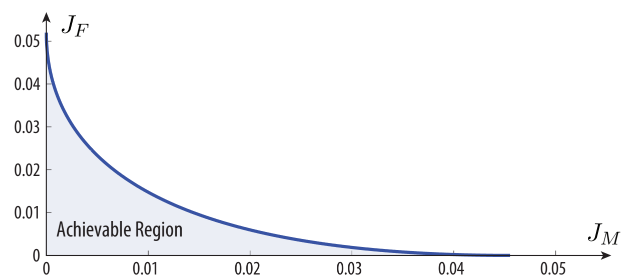

Fig. 1. Achievable region when $\varepsilon = 0.35$ given by the lower-bound in Theorem 1.1 and upper-bounds in Theorem 2.1. The upper-bound matches the lower-bound completely.

3. PROOF

The main idea of our proof is the following: Consider any $J_1 > 0$. Suppose there is a test ϕ such that $J_F(\phi) \geq J_1$. We would like to prove that $J_M(\phi) \leq J_2$. Consider any $\delta > 0$. We construct a sequence of events $\{A_n\}$ that satisfies the following:

$$\lim_{n\to\infty} -\frac{m}{n^2}\log(\mathsf{P}_\pi(A_n)) \leq J_1 - \delta; \tag{10}$$

For any $\boldsymbol{z}_1^n$ satisfying $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\} \subseteq A_n$, we have:

$$\sup_{\mu\in\Pi_m} \frac{\mu^n}{\pi^n}(\boldsymbol{z}_1^n) \geq \exp\{-\frac{n^2}{m}(J_2 - J_1 + \delta)\}. \tag{11}$$

Since $J_F(\phi) \geq J_1$, we conclude from (10) that $\mathsf{P}_\pi(\phi = 0|A_n) = 1 - o(1)$. Then the upper-bound on generalized error exponent can be obtained from (11)

$$P_M(\phi_n) \geq P_\pi(A_n)\mathsf{P}_\pi(\phi = 0|A_n)\exp\{-\frac{n^2}{m}(J_2 - J_1 + \delta)\}$$

Proof. We only prove for the case where $\varepsilon \geq 0.5$. The proof for the case where $\varepsilon < 0.5$ is essentially the same but more tedious. Since the case where $\tau = 0$ has been proved in Theorem 1.2, we only consider $\tau > 0$. For simplicity of exposition, denote

$$f_1(\tau,\varepsilon) = \sup_{\theta\geq 0}\{\theta\tau - \tfrac{1}{2}(e^{2\theta}-1-2\theta)\},$$
$$f_2(\tau,\varepsilon) = \sup_{\theta\geq 0}\{\theta(\kappa(\varepsilon)-\tau) - \tfrac{1}{2}(e^{-2\theta}-1+2\theta)(1+\kappa(\varepsilon))\}.$$

Define the event

$$A_n = \{K_n \leq n-(1+\tau-\delta)\frac{n^2}{m}\}.$$

As a consequence of Theorem 1.1, we obtain

Lemma 3.1.

$$\lim_{n\to\infty} -\frac{m}{n^2}\log \mathsf{P}_\pi(A_n) = f_1(\tau-\delta,\varepsilon).$$

Let K_m denote the collection of all subsets of $[m]$ whose cardinality is $\lfloor m(1-\varepsilon)\rfloor$. For each $\mathcal{U}\in K_m$, define the distribution

$$\mu_{\mathcal{U},j} = \begin{cases} \frac{1}{\lfloor m(1-\varepsilon)\rfloor}, & j\in\mathcal{U}; \\ 0, & j\in[m]\setminus\mathcal{U}. \end{cases} \tag{12}$$

Consider the mixture $\bar{\mu}^n = \frac{1}{|K_m|}\sum_{\mathcal{U}\in K_m}\mu^n_{\mathcal{U}}$. The following lower-bound on the average likelihood ratio $\bar{\mu}^n/\pi^n$ holds:

Lemma 3.2. *For any sequence $\boldsymbol{z}_1^n = \{z_1,\ldots,z_n\}$ such that $\{\boldsymbol{Z}_1^n = \boldsymbol{z}_1^n\}\subseteq A_n$,*

$$\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) \geq \exp\{-\tfrac{1}{2}\frac{n^2}{m}[\frac{\varepsilon}{1-\varepsilon}+\log(1-\varepsilon)(1+\tau-\delta)]+O(\frac{n^3}{m^2})\}.$$

Consider any test ϕ such that $J_F(\phi)\geq f_1(\tau,\varepsilon)$. Comparing this with Lemma 3.1, we conclude that

$$\mathsf{P}_\pi\{\phi_n = 0|A_n\} = 1-o(1).$$

We then have

$$\begin{aligned} & P_M(\phi_n) \\ \geq\ & \sup_{\mu\in\Pi_m}\mathsf{P}_\mu\{\phi_n=0|A_n\}\mathsf{P}_\pi(A_n) \\ \geq\ & \mathsf{P}_{\bar{\mu}}\{\phi_n=0|A_n\}\mathsf{P}_\pi(A_n) \\ \geq\ & \mathsf{P}_\pi\{\phi_n=0|A_n\}\mathsf{P}_\pi(A_n) \\ & \times\exp\{-\tfrac{1}{2}\frac{n^2}{m}[\frac{\varepsilon}{1-\varepsilon}+\log(1-\varepsilon)(1+\tau-\delta)]+O(\frac{n^3}{m^2})\}. \end{aligned}$$

Consequently,

$$\begin{aligned} & J_M(\phi) \\ \leq\ & \tfrac{1}{2}\frac{\varepsilon}{1-\varepsilon}+\tfrac{1}{2}\log(1-\varepsilon)(1+\tau-\delta)+f_1(\tau-\delta,\varepsilon) \\ =\ & \tfrac{1}{2}[\kappa(\varepsilon)+\log(1+\kappa(\varepsilon))(1+\tau-\delta)] \\ & +\tfrac{1}{2}[(\tau-\delta)\log(1+\tau-\delta)-\tau+\delta+\log(1+\tau-\delta)] \\ =\ & f_2(\tau,\varepsilon)+h(\delta). \end{aligned} \tag{13}$$

where

$$h(\delta) = \tfrac{1}{2}[-\delta\log(1+\kappa(\varepsilon))+(1+\tau)\log(1-\frac{\delta}{1+\tau})-\delta\log(1+\tau-\delta)+\delta].$$

Note that $\lim_{\delta\to\infty}h(\delta)=0$. Since the inequality (13) holds for any $\delta>0$, we conclude that $J_M(\phi)\leq f_2(\tau,\varepsilon)$. □

Proof of Lemma 3.2. Let $\mathcal{S} := \{j\ :\ j \text{ appears in } \boldsymbol{z}_1^n\}$. Let $s=|\mathcal{S}|$. Since $\{\boldsymbol{Z}_1^n=\boldsymbol{z}_1^n\}\subseteq A_n$, and $2(s-H_1)+H_1\leq n$, we obtain

$$s\leq n-\tfrac{1}{2}(1+\tau-\delta). \tag{14}$$

The likelihood ratio $\mu^n_{\mathcal{U}}/\pi^n$ is given by

$$\frac{\mu^n_{\mathcal{U}}}{\pi^n}(\boldsymbol{z}_1^n) = (\frac{m}{\lfloor m(1-\varepsilon)\rfloor})^n\mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}}.$$

Consequently,

$$\frac{\bar{\mu}^n}{\pi^n}(\boldsymbol{z}_1^n) = (\frac{m}{\lfloor m(1-\varepsilon)\rfloor})^n(\frac{1}{|K_m|}\sum_{\mathcal{U}\in K_m}\mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}}), \tag{15}$$

where

$$\frac{1}{|K_m|}\sum_{\mathcal{U}\in K_m}\mathbb{I}_{\mathcal{S}\subseteq\mathcal{U}} = \binom{m-s}{\lfloor m(1-\varepsilon)\rfloor - s}\Big/\binom{m}{\lfloor m(1-\varepsilon)\rfloor}.$$

Stirling's formula leads the following asymptotic approximation of right hand side:

$$\binom{m-s}{\lfloor m(1-\varepsilon)\rfloor-s}\Big/\binom{m}{\lfloor m(1-\varepsilon)\rfloor} = (1-\varepsilon)^s\exp\{-\tfrac{1}{2}\frac{s^2}{m}\frac{\varepsilon}{1-\varepsilon}+O(\frac{s^3}{m^2})\}.$$

Substituting this and (14) into (15), we obtain the claim of the lemma. □

4. CONCLUSION

The hardness result presented in this paper gives a tight upper-bound on the achievable generalized error exponents for goodness of fit test. The performance of the coincidence-based test achieves this upper-bound, and is optimal in terms of generalized error exponents. Future research directions include:

(i) There are other tests, such as the weighted coincidence-based test introduced in [1] that achieve the same generalized error exponent. To compare these tests, one needs to look at finer criteria, such as sharp large deviation analysis [3].

(ii) For goodness of fit problem with large number of samples (m=O(n)), the converse result established in [4] is an upper-bound on the (classical) error exponent of missed detection. It is possible that the technique used in this paper is also applicable to that case to give a full converse result, i.e., upper-bounds on both error exponent of missed detection and false alarm.

5. REFERENCES

[1] 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 (ICASSP 2012).

[2] L. Paninski, "A coincidence-based test for uniformity given very sparsely sampled discrete data," *IEEE Trans. Inf. Theory*, vol. 54, no. 10, pp. 4750 – 4755, Oct. 2008.

[3] R. R. Bahadur and R. R. Rao, "On deviations of the sample mean," *Ann. Math. Statist.*, vol. 31, no. 4, pp. 1015 – 1027, 1960.

[4] A. R. Barron, "Uniformly powerful goodness of fit tests," *Ann. Statist.*, vol. 17, no. 1, pp. 107 – 124, 1989.