The Sample Complexity of Simple Binary Hypothesis Testing

Ankit Pensia ANKITP@IBM.COM
IBM Research
Varun Jog VJ270@CAM.AC.UK
University of Cambridge
Po-Ling Loh PLL28@CAM.AC.UK
University of Cambridge

Editors: Shipra Agrawal and Aaron Roth

Hypothesis testing is a fundamental problem in statistical inference that seeks the most likely hypothesis corresponding to a given set of observations. The simplest formulation of hypothesis testing, which is also the focus of this paper, is *simple binary hypothesis testing*. Here, the two hypotheses correspond to distributions p and q over a domain $\mathcal{X}$, and the set of observations comprises n i.i.d. samples $X_1, \dots, X_n$ from $\theta \in \{p, q\}$. The goal is to identify which distribution generated the samples; i.e., to produce $\widehat{\theta} := \widehat{\theta}(X_1, \dots, X_n)$ so that $\widehat{\theta} = \theta$ with high probability.

Simple binary hypothesis testing is a crucial building block in statistical inference procedures. Consequently, much research has analyzed the best algorithms and their performance (Lehmann et al., 1986). The famous Neyman–Pearson lemma (Neyman and Pearson, 1933) provides the optimal procedure for $\widehat{\theta}$, and subsequent works have completely characterized its error probability in two regimes: the single-sample setting ($n = 1$) and the infinite-sample setting ($n \to \infty$). While the single-sample regime is relatively straightforward, much historical work in statistics and information theory has focused on the infinite-sample (or asymptotic) regime, where the asymptotic error admits particularly neat expressions in terms of information-theoretic divergences between p and q (see the textbooks Cover and Thomas (2006); Polyanskiy and Wu (2023).

Although asymptotic results provide crucial insight into the problem structure, they offer no concrete guarantees for finite samples. In particular, asymptotic bounds cannot satisfactorily answer the question of how many samples are needed (i.e., what is the sample complexity) to solve hypothesis testing with a desired level of accuracy. This limits their applicability in practice, as well as in learning theory research, where sample complexity bounds are paramount.

In the context of simple binary hypothesis testing, an algorithm can make two types of errors: (i) $\mathbb{P}(\widehat{\theta} = q | \theta = p)$, termed type-I error, and (ii) $\mathbb{P}(\widehat{\theta} = p | \theta = q)$, termed type-II error. These two types of errors may have different operational consequences in different situations, e.g., false positives vs. false negatives of a lethal illness, where the cost incurred by the former is significantly smaller than that of the latter. A natural way to combine these metrics is by considering a weighted sum of the errors, which is equivalent to the well-studied Bayesian formulation. We also consider a prior-free version, where these two errors are analyzed separately.

While some prior work has established non-asymptotic bounds on the error probability in simple binary hypothesis testing, we observe that the sample complexity perspective has remained largely unaddressed (cf. Strassen (1962); Polyanskiy et al. (2010); Bar-Yossef (2002)). Our main contribution is to fill this gap by developing tight results for the sample complexity of simple binary hypothesis testing for both the Bayesian formulation and the prior-free formulation.[1]

1. Extended abstract. Full version appears as [arXiv:2403.16981,v2].

References

Z. Bar-Yossef. *The complexity of massive data set computations*. PhD thesis, University of California, Berkeley, 2002.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

E. L. Lehmann, J. P. Romano, and G. Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

J. Neyman and E. S. Pearson. On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231:289–337, 1933. ISSN 0264-3952.

Y. Polyanskiy and Y. Wu. *Information Theory: From Coding to Learning,*. Cambridge University Press, 2023.

Y. Polyanskiy, H. V. Poor, and S. Verdú. Channel coding rate in the finite blocklength regime. *IEEE Transactions on Information Theory*, 56(5):2307–2359, 2010.

V. Strassen. Asymptotische abschatzugen in Shannon's informationstheorie. In *Transactions of the Third Prague Conference on Information Theory etc, 1962. Czechoslovak Academy of Sciences, Prague*, pages 689–723, 1962.