The Sample Complexity of Distributed Simple Binary Hypothesis Testing under Information Constraints

Hadi Kazemi HK569@CAM.AC.UK
University of Cambridge
Ankit Pensia ANKITP@BERKELEY.EDU
Simons Institute, UC Berkeley
Varun Jog VJ270@CAM.AC.UK
University of Cambridge

Editors: Nika Haghtalab and Ankur Moitra

Keywords: Sample complexity, hypothesis testing, data-processing inequality

1. Introduction

Hypothesis testing is one of the central problems in statistics that concerns testing different hypotheses based on observed data. The most basic version of this problem is simple binary hypothesis testing, where we wish to decide between two hypotheses H_0 and H_1. Under hypothesis H_0 (resp. H_1) we observe i.i.d. samples from distribution Q (resp. P), where P and Q are known. The Bayes formulation of simple binary hypothesis testing, which is the focus of this paper, assumes that the true hypothesis is chosen randomly according to a prior Bernoulli distribution $\mathrm{Ber}(\pi)$. In this paper we study the distributed version of hypothesis testing. Suppose each of n agents, denoted by $s_1, \ldots, s_n$, receives a single sample denoted by X_i for agent s_i. Agent s_i transmits Y_i to a central server, and the central server performs a hypothesis test using $Y_1, \ldots, Y_n$. The transformation of X_i to Y_i is subject to information constraints that are captured by a set of channels $\mathcal{T}$ (Markov kernels) from $\mathcal{X}$, the support of the X_is, to $\mathcal{Y}$, the support of the Y_is; i.e., agent s_i must pick a communication channel $T^i \in \mathcal{T}$. For example, $\mathcal{T}$ could be the set of channels with output size D, or the set of all ϵ-differentially private channels. Agents may choose in the following manners: *(i) identical chanels:* all channels are the same, *(ii) non-identical channels:* different agents may choose different channels *(iii) interactive setting:* each agent s_i chooses a channel after observing prior messages $Y_1, Y_2, \ldots, Y_{i-1}$. This paper resolves two open problems from Pensia et al. (2024) concerning the sample complexity of distributed simple binary hypothesis testing under information constraints.

Question 1: Does interactivity help significantly reduce (i.e., by more than constant factors) the sample complexity of distributed simple binary hypothesis testing?

Question 2: Are communication constraints benign, i.e., do they increase the sample complexity by at most logarithmic factors compared to the unconstrained setting, in all regimes of interest of the parameters? In this paper, we show that sequential interaction reduces the sample complexity by a factor of 4, at best. We answer the second question in the affirmative by deriving optimally tight bounds for the sample complexity. Our main technical contributions are: (i) a one-shot lower bound on the Bayes error in simple binary hypothesis testing that satisfies a crucial tensorisation property; (ii) a streamlined proof of the formula for the sample complexity of simple binary hypothesis testing without constraints, first established in Pensia et al. (2024); and (iii) a reverse data-processing inequality for Hellinger-λ divergences, generalising the results from Bhatt et al. (2021) and Pensia et al. (2023).[1]

1. Extended abstract. Full version appears as [arXiv:2506.13686]

References

A. Bhatt, B. Nazer, O. Ordentlich, and Y. Polyanskiy. Information-distilling quantizers. *IEEE Transactions on Information Theory*, 67(4):2472–2487, 2021.

A. Pensia, V. Jog, and P. Loh. Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities. *IEEE Transactions on Information Theory*, 2023.

A. Pensia, V. Jog, and P. Loh. The sample complexity of simple binary hypothesis testing. In *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247, pages 4205–4206. PMLR, 2024.