Simple Binary Hypothesis Testing under Local Differential Privacy and Communication Constraints

Ankit Pensia ANKITP@CS.WISC.EDU
University of Wisconsin-Madison

Amir R. Asadi ASADI@STATSLAB.CAM.AC.UK
Varun Jog VJ270@CAM.AC.UK
Po-Ling Loh PLL28@CAM.AC.UK
University of Cambridge

Editors: Gergely Neu and Lorenzo Rosasco

We study simple binary hypothesis testing under local differential privacy (LDP) and communication constraints, focusing on simple binary hypothesis testing, a fundamental problem in statistical estimation. We will also consider the more realistic setting of LDP constraints in tandem with communication constraints. The case with communication constraints alone was addressed by Pensia, Jog, and Loh (2022).

Let p and q be two distributions over a finite domain $\mathcal{X}$, and let $X_1, \dots, X_n \in \mathcal{X}^n$ be i.i.d. samples drawn from either p or q. The goal of simple hypothesis testing is to identify (with high probability) whether the samples were drawn from p or q. This problem has been extensively studied in both asymptotic and nonasymptotic settings (Neyman and Pearson, 1933; Wald, 1945; Cam, 1986). In the context of LDP, the statistician no longer has access to the original samples $X_1, \dots, X_n$, but only their privatized counterparts $Y_1, \dots, Y_n \in \mathcal{Y}^n$, for some set $\mathcal{Y}$. Each X_i is transformed to Y_i via a private channel $\mathbf{T}_i$, which is simply a probability kernel specifying $\mathbf{T}_i(y, x) = \mathbb{P}(Y_i = y | X_i = x)$. We characterize the sample complexity $\mathrm{n}^*(p, q, \epsilon)$ of this task, i.e., the smallest n for which there exists a protocol whose error probability for hypothesis testing is less than a constant, say, 0.1.

Our results are either minimax optimal or instance optimal: the former hold for the set of distribution pairs with prescribed Hellinger divergence and total variation distance, whereas the latter hold for specific distribution pairs. For the sample complexity of simple hypothesis testing under pure LDP constraints, we establish instance-optimal bounds for distributions with binary support; minimax-optimal bounds for general distributions; and (approximately) instance-optimal, computationally efficient algorithms for general distributions. Under both privacy and communication constraints, we develop instance-optimal, computationally efficient algorithms that achieve minimal sample complexity (up to universal constants). Our results on instance-optimal algorithms hinge on identifying the extreme points of the joint range set of two distributions p and q, defined as $\mathcal{A} := \{(\mathbf{T}p, \mathbf{T}q) | \mathbf{T} \in \mathcal{C}\}$, where $\mathcal{C}$ is the set of channels characterizing the constraints.

Existing results on simple binary hypothesis testing under LDP constraints have focused on the high-privacy regime of $\epsilon \in (0, c)$, for a constant $c > 0$. They do not inform sample complexity when $1 \ll \epsilon < \infty$. This is not an artifact of analysis: the optimal tests in the low- and high-privacy regimes are fundamentally different. The large-ϵ regime has been increasingly used in practice, due to privacy amplification provided by shuffling (Cheu et al., 2019; Bittau et al., 2017; Feldman et al., 2021). Our paper makes progress on the computational and statistical fronts in the large-ϵ regime.[1]

1. Extended abstract. Full version appears as arXiv:2301.03566, v2.

References

A. Bittau, Ú. Erlingsson, P. Maniatis, I. Mironov, A. Raghunathan, D. Lie, M. Rudominer, U. Kode, J. Tinnes, and B. Seefeld. Prochlo: Strong privacy for analytics in the crowd. In *Proc. of the 26th Symposium on Operating Systems Principles*, 2017.

L. L. Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer New York, New York, NY, 1986. ISBN 978-1-4612-9369-9.

A. Cheu, A. Smith, J. Ullman, D. Zeber, and M. Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology – EUROCRYPT 2019*, 2019.

V. Feldman, A. McMillan, and K. Talwar. Hiding among the clones: A simple and nearly optimal analysis of privacy amplification by shuffling. In *Proc. 62nd IEEE Symposium on Foundations of Computer Science (FOCS)*, 2021.

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.

A. Pensia, V. Jog, and P. Loh. Communication-constrained hypothesis testing: Optimality, robustness, and reverse data processing inequalities. *CoRR*, arXiv:2206.02765, 2022.

A. Wald. Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2): 117–186, 1945. ISSN 0003-4851. doi: 10.1214/aoms/1177731118.