PAPRIKA: Private Online False Discovery Rate Control

Wanrong Zhang Gautam Kamath Rachel Cummings

Introduction. In hypothesis testing, a *false discovery* occurs when a hypothesis is incorrectly rejected due to noise in the sample. When adaptively testing multiple hypotheses, the probability of a false discovery increases as more tests are performed. Thus the problem of *False Discovery Rate (FDR) control* is to find a procedure for testing multiple hypotheses that accounts for this effect in determining the set of hypotheses to reject. The goal is to minimize the number (or fraction) of false discoveries, while maintaining a high true positive rate (i.e., correct discoveries).

In this work (full version available at: https://arxiv.org/abs/2002.12321), we study False Discovery Rate (FDR) control in multiple hypothesis testing under the constraint of differential privacy for the sample. Unlike previous work in this direction focusing on the offline setting, we focus on the *online setting*. A data analyst receives a stream of hypotheses (equivalently, a stream of p values $p_1, p_2, \ldots$) on the database D. The analyst picks a threshold α_t at each time t to reject the hypothesis if $p_t \leq \alpha_t$; the threshold can depend on previous discoveries, and rejection must be decided before the next hypothesis arrives. The error metric FDR is formally defined as $\text{FDR} = \mathbb{E}\left[\text{FDP}\right] = \mathbb{E}\left[\frac{|\mathcal{H}^0 \cap \mathcal{R}|}{|\mathcal{R}|}\right]$.

Main Results. We provide a novel algorithm for private online false discovery rate control, PAPRIKA, which takes in a stream of p-values, a target FDR level, and privacy parameters ε and δ, and it outputs discoveries that can control the FDR at a certain level at any time point. It starts with the state-of-the-art online FDR control algorithm SAFFRON [2], using SPARSEVECTOR to ensure privacy of the rejection set. However, the combination of these tools is far from immediate, and several algorithmic innovations are required, including: dynamic thresholds in SPARSEVECTOR to accommodate the alpha-investing rule, adding noise that scales with the multiplicative sensitivity of p-values to reduce the noise required for privacy, shifting the SparseVector threshold to accommodate FDR as a novel accuracy metric, and the candidacy indicator step which cannot be done privately and requires modifications to the wealth updates. We resolve this by using a similar wealth updating rule as in another non-private algorithm LORD++ [1]. PAPRIKA provides unconditional differential privacy guarantees and satisfies the FDR control guarantees. We also provide experimental results to demonstrate the efficacy of our algorithms in a variety of data environments.

Theorem 1. *For any stream of p-values* $\{p_1, p_2, \ldots\}$, PAPRIKA *is* (ε, δ)*-differentially private.*

Theorem 2 (FDR control (informal))**.** *If the null p-values are independent of each other and of the non-null p-values, then* PAPRIKA *controls* $FDR(t) \leq \alpha + \delta t$ *for all* $t \in \mathbb{N}$.
It can control a slightly weaker notion of FDR if the null p-values are adaptive (under certain conditions).

The FDR bounds provided by PAPRIKA are weaker by an additive δt, relative to the non-private guarantees. In most differential privacy applications, δ is typically required to be cryptographically small, so this additional term should have a minuscule effect on the FDR. We note that ε plays a role in the analysis of Theorem 2, although it does not appear in FDR bounds.

References

[1] A. Ramdas, F. Yang, M. J. Wainwright, and M. I. Jordan. Online control of the false discovery rate with decaying memory. In *Advances In Neural Information Processing Systems*, pages 5650–5659, 2017.

[2] A. Ramdas, T. Zrnic, M. Wainwright, and M. Jordan. SAFFRON: an adaptive algorithm for online control of the false discovery rate. In *International Conference on Machine Learning*, pages 4286–4294, 2018.