Efficient and Optimal Learning of Discrete Distributions with Person-Level Privacy

Abstract

We study learning a discrete distribution under person-level differential privacy in the batched setting. We are given n users, each contributing a batch of m samples drawn i.i.d. from an unknown distribution μ over $[k]$, and seek an estimate of μ up to total-variation error α. We give polynomial-time algorithms under both pure and approximate differential privacy, with sample-complexity bounds matching the information-theoretic optima up to logarithmic factors. Under pure $(\varepsilon, 0)$-DP, we achieve sample complexity

$$n = \tilde{O}\left(\frac{k}{\alpha^2 m} + \frac{k}{\alpha\sqrt{m}\,\varepsilon} + \frac{k}{\varepsilon}\right).$$

In particular, the pure-DP result gives the first polynomial-time pure-DP algorithm for this problem with near-optimal sample complexity.

Our approach instantiates the robustness-to-privacy framework with a new robust core for discrete batches. The main ingredient is a polynomial-time single-shot sum-of-squares estimator for learning from untrusted batches. At its core is a succinct covariance certificate, given by a trace-bounded diagonal majorizer for the corrected subset-variance discrepancy. This certificate preserves the mean-dependent variance structure of multinomial batch averages, yields the near-optimal robust rate $O(\eta\sqrt{\log(1/\eta)/m})$, allowing it to serve as the score object required by the privacy reduction.