Private and polynomial time algorithms for learning Gaussians and beyond: Extended Abstract

Hassan Ashtiani
McMaster University

Christopher Liaw
Google

Editors: Po-Ling Loh and Maxim Raginsky

Learning a multivariate Gaussian distribution with respect to the total variation (TV) distance is one of the most basic tasks in statistical estimation. It is folklore that simply taking the empirical mean and covariance matrix achieves the optimal sample complexity of $\Theta(d^2/\alpha^2)$ for learning unbounded Gaussians in $\mathbb{R}^d$ up to TV distance $\alpha$. However, this simple approach can compromise the privacy of the individuals/entities whose data has been used in the estimation procedure. A central question that we aim to address is whether unbounded Gaussians can be learned privately, in the sense of approximate $(\epsilon, \delta)$-differential privacy ($((\epsilon, \delta)$-DP) and efficiently.

Motivated by this question, we present a fairly general and simple framework for reducing $(\epsilon, \delta)$-DP statistical estimation to its non-private counterpart. At a high-level, our framework operates as follows. Suppose that we have a non-private estimator $A$. Given a set of samples, we split the samples into disjoint subsets and run $A$ on each subset. We privately check whether most of the (non-private) outputs of $A$ are clustered together. If so, we aggregate the points that are close to the cluster using a weighted average and release a noisy version of this weighted average.

As the main application of this framework, we give a polynomial time and $(\epsilon, \delta)$-DP algorithm for learning (unrestricted) Gaussian distributions in $\mathbb{R}^d$. The sample complexity of our approach for learning the Gaussian up to total variation distance $\alpha$ is $\tilde{O}(d^2/\alpha^2 + d^2\sqrt{\ln(1/\delta)/\alpha\epsilon} + d\ln(1/\delta)/\alpha\epsilon)$ matching (up to logarithmic factors) the best known information-theoretic (non-efficient) sample complexity upper bound due to Aden-Ali et al. (2021). In an independent work and using a completely different technique, Kamath et al. (2021) proved a similar result but with higher dependence of $O(d^{5/2})$ on the dimension $d$.

A technical challenge in our application to learning Gaussians is to design a “masking” mechanism $B$ such that if two positive-definite matrices $\Sigma_1, \Sigma_2$ are close in Mahalanobis distance then $B(\Sigma_1), B(\Sigma_2)$ are indistinguishable. One cannot simply add Gaussian noise to the input matrices since this requires a bound on the Frobenius distance. In general, there is no finite bound on the ratio between the Mahalanobis distance and the Frobenius distance (the former adapts to the geometry of the input matrices whereas the latter does not). Thus, adding Gaussian noise cannot work. However, a key observation is that, in some sense, the Mahalanobis distance still has some resemblance to the Frobenius distance. With this observation in hand, we show that adding covariance-rescaled Gaussian noise does allow us to ensure that the input is indistinguishable.

As a second application of our framework, we provide the first polynomial time $(\epsilon, \delta)$-DP algorithm for robust learning of (unrestricted) Gaussians whose sample complexity is $\tilde{O}(d^{3.5})$ in terms of the dimension. In another independent work, Kothari et al. (2021) also provided a polynomial time $(\epsilon, \delta)$-DP algorithm for robust learning of Gaussians with sample complexity $O(d^8)$.

Acknowledgments

HA is affiliated with Vector Institute and was supported by an NSERC Discovery grant. CL was supported by an NSERC fellowship and the Vector Institute. Part of this work was done while CL was a postdoc at the University of Toronto.

References

