Adaptive Data Analysis refers to the reuse of data to perform analyses suggested by the outcomes of previously computed statistics on the same data. In this work, we initiate a principled study of how the generalization properties of approximate differential privacy can be used to perform adaptive hypothesis testing. This substantially extends the existing connection between differential privacy and max-information, which previously was only known to hold for pure differential privacy. It also extends our understanding of max-information as a partially unifying measure controlling the generalization properties of adaptive data analyses.

Adaptive Data Analysis

- A lot of existing theory assumes tests are selected independently of the data.
- In practice, data analysis is inherently interactive, where experiments may depend on previous outcomes from the same dataset.
- Question: How can we provide statistically valid answers to adaptively chosen analyses?
- Answer: Limit the information learned about the dataset. [DFH+15a]
- Part of a line of work initiated by [DFH+15a, DFH+15b, HU14].

Post-Selection Hypothesis Testing

- Hypothesis test: Defined by a null hypothesis $H_0$ and a test statistic $t$.
- Purpose: Reject $H_0$ if the data $X$ is not likely to have been generated from some distribution $Q^n$ such that $Q \in H_0$.
- Significance level of $\alpha = \Pr_{X \sim Q^n}[t(X) = \text{Reject}] \leq \alpha$.
- Assumes choice of $t$ is independent of the data $X$.
- Goal: For an adaptively chosen test $t_{A(X)}$, we want to bound $\Pr_{X \sim Q^n}[t_{A(X)}(X) = \text{Reject}]$ for $Q \in H_0$.
- Problem: $t_{A(X)}$ can be tailored specifically to $X$.

Max-Information [DFH+15b]

- An algorithm $A$ with bounded max-info allows the analyst to treat the output $A(X)$ as if it is independent of data $X$ up to a factor. $t^\beta_{\alpha}(A(X)) \leq k \iff \Pr_{X \sim Q^n}[t_{A(X)}(X) = \text{Reject}] \leq \beta$.
- Differentiate between general and product distributions: $t^\beta_{\alpha}(A(n)) = \sup_{X \in S} I^\beta_{\alpha}(X; A(X))$.
- [RRST.16]: If $t^\beta_{\alpha}(A(n)) \leq k$, then for $\gamma(\alpha) = (k^2 + 1)^{1/2}$.

Differential Privacy [DMNS06]

- A randomized algorithm $A: D^n \to T$ is $(\epsilon, \delta)$-differentially private if for all neighboring data sets $x, y \in D^n$, i.e., $\text{dist}(x, y) = 1$, and for all sets of outcomes $S \subseteq T$, we have $P(A(x) \in S) \leq e^\epsilon P(A(y) \in S) + \delta$.
- If $\delta = 0$, we say pure DP. If $\delta > 0$, we say approximate DP.

Technical Contributions

- Previous results [DFH+15a]: If $A: D^n \to T$ is $(\epsilon, \delta)$-DP, then for $\beta > 0$, we have $I^\beta_{\alpha,0}(A(n)) \leq O(\epsilon^2 n)$.
- Positive Result: If $A: D^n \to T$ is $(\epsilon, \delta)$-DP, for $\beta \approx O(n, \sqrt{\delta}/\epsilon)$, we have $I^\beta_{\alpha,0}(A(n)) = O(\epsilon^2 n + n, \sqrt{\delta}/\epsilon)$.
- Approx. DP $\Rightarrow$ Max-Information

Related Publications


Supported in part by grants from the Sloan foundation and NSF grants: CNS-1253345, CNS-1513694, IIS-1447700.