Max-Information, Differential Privacy, and Post-Selection Hypothesis Testing

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar

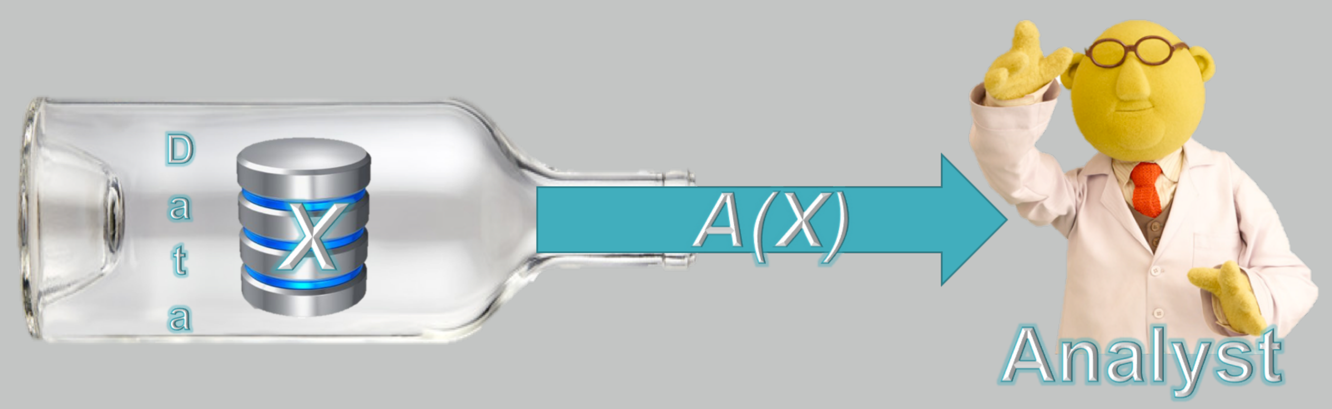
Penn University of Pennsylvania

PennState 1855

Overall Goal: Maintain statistical validity in adaptive data analysis.

Adaptive Data Analysis

- Part of a line of work initiated by [DFH+15a, DFH+15b, HU14].
- In practice, data analysis is inherently interactive, where experiments may depend on previous outcomes from the same dataset.
- To allow the analyst to reuse the dataset for multiple experiments, we want to restrict the amount of information learned about the data so that later experiments are *nearly* independent of the data.



False Discovery

- We want to design valid hypothesis tests where probability of a false discovery $< \alpha$.
- Design a test t and use a p-value to determine if model H_0 is likely given the data
$$p(a) = \mathbb{P}_{X\sim H_0}(t(X) > a)$$
- Note that $p(t(X)) \sim \text{Unif}[0,1]$. Rejecting H_0 if $p(t(X)) < \alpha$ ensures false discovery is at most α.
- Framework crucially relies on test being chosen independent of the data.
- Has led to false discovery rates $> \alpha$.

Valid p-Value Correction

$\gamma : [0,1] \to [0,1]$ is a valid p-value correction for selection procedure A if $\forall \alpha$ the following procedure has false discovery $\leq \alpha$:

(1) Select test $t \leftarrow A(X)$

(2) Reject H_0 if the p-value $p(t(X)) \leq \gamma(\alpha)$

Max-Information [DFH+15b]

- An algorithm A with bounded max-info allows the analyst to treat $A(X)$ as if it is independent of data X up to a factor.
$$I_\infty^\beta(A(X), X) = \log\left(\sup_O \frac{\mathbb{P}((A(X), X) \in O) - \beta}{\mathbb{P}((A(X)\otimes X) \in O)}\right)$$
- Differentiate between product and general distributions
$$I_\infty^\beta(A; n) = \sup_{\mathcal{S}: X\sim\mathcal{S}} I_\infty^\beta(A(X), X)$$
$$I_{\infty,\mathcal{P}}^\beta(A; n) = \sup_{\mathcal{P}: X\sim\mathcal{P}^n} I_\infty^\beta(A(X), X)$$
- $I_{\infty,\mathcal{P}}^\beta(A; n) \leq k$, leads to a p-value correction:
$$\gamma(\alpha) = (\alpha - \beta)/2^k$$

Algorithms with Bounded Max Info [DFH+15b]

Algorithms $A : D^n \to \mathcal{Y}$ with bounded max-info include:

(1) Pure ϵ-differentially private algorithms
$$I_{\infty,\mathcal{P}}^\beta(A; n) \leq \epsilon\sqrt{n\log(1/\beta)}$$
$$I_\infty^0(A; n) \leq \epsilon n$$

(2) Bounded description length algorithms.
$$I_\infty^\beta(A; n) \leq \log(|\mathcal{Y}|/\beta)$$

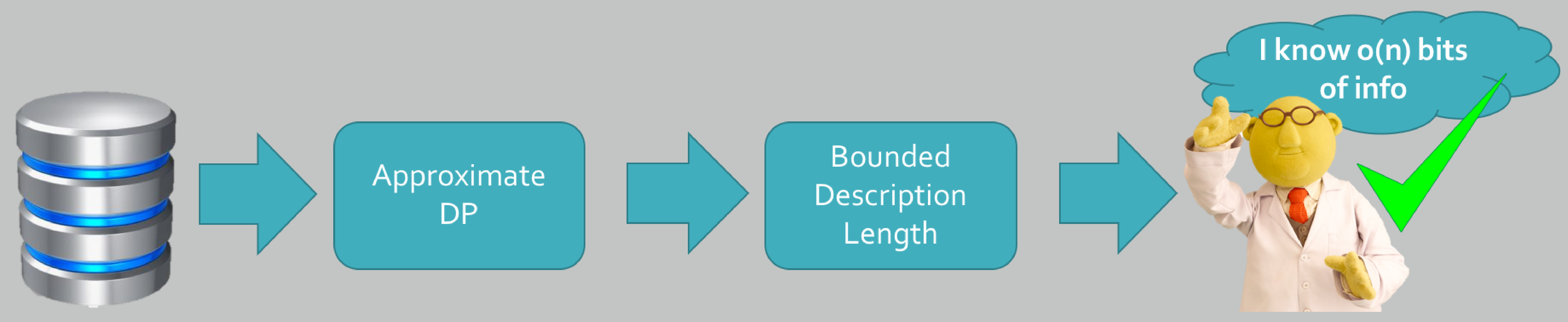
What About Max-Info for Approximate Differential Privacy?

- Best known algorithms for adaptive data analysis are approximate DP.
- Can we get better Max-Info bounds for (ϵ, δ)-DP.
- Huge improvement in using approximate differential privacy in composition: using T many ϵ-DP algorithms leads to ϵT-DP but also $(\epsilon\sqrt{T\log(1/\delta)}, \delta)$ -DP.

Positive Result

If $A : D^n \to \mathcal{Y}$ is (ϵ, δ)-DP then
$$\underbrace{I_{\infty,\mathcal{P}}^\beta(A; n)}_{\text{Product Distributions}} \leq \left(\epsilon^2 + \sqrt{\epsilon\delta}\right) n, \qquad \beta \leq n\sqrt{\frac{\delta}{\epsilon}}$$

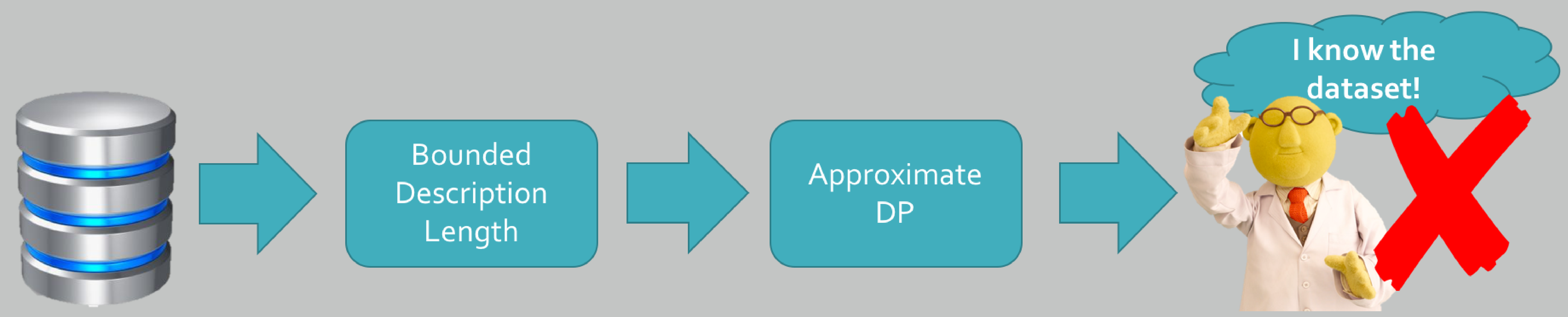
- Nearly gives the tight generalization bounds of DP algorithms for low sensitive queries from [BNS+16], but does NOT apply to p-values.
- [RZ16] also give a method to correct p-values based on mutual info but we can get an improved correction factor via Max-Info.



Negative Result

There exists and (ϵ, δ)-DP algorithm such that
$$\underbrace{I_\infty^\beta(A; n)}_{\text{General Distributions}} \geq n - \log(1/\delta)\log(n)/\epsilon$$

- We know that Max-Info composes and so pure DP and bounded description length algorithms can be used in any order.
- **Ordering matters**: we prove the negative result by showing that composing a bounded description length algorithm followed by an approx-DP algorithms outputs the full dataset w.h.p.



Acknowledgements and References

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

[BNS+16] Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. In *STOC*, 2016.

[DFH+15a] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. In *NIPS*. 2015.

[DFH+15b] Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Leon Roth. In *STOC*, 2015.

[HU14] Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *FOCS*, 2014.

[RZ16] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. In *AISTATS*, 2016.