Sparse Vector

Katrina Ligett

key idea

What if we only want to answer queries whose answers fall *above a certain threshold*, or *in a certain range*?

What if we could cheaply check whether we can answer a new query using information we’ve gathered from past queries, and only “pay” for the new query if we can’t already approximate its answer well?

…sounds like reportNoisyMax

If all the queries were defined in advance, could use reportNoisyMax or Exponential Mechanism

If queries *adaptive*, more subtle

aboveThreshold

a simpler problem first: given adaptive sequence of queries and a threshold,

- return “below” for queries (approximately) below the threshold
- return “above” and halt once you reach a query that’s (approximately) above the threshold

aboveThreshold

Algorithm 1 Input is a private database D, an adaptively chosen stream of sensitivity 1 queries $f_1, \ldots,$ and a threshold T. Output is a stream of responses $a_1, \ldots$

AboveThreshold$(D, \{f_i\}, T, \epsilon)$

  Let $\hat{T} = T + \text{Lap}\left(\frac{2}{\epsilon}\right)$.
  for Each query i **do**
    Let $\nu_i = \text{Lap}(\frac{4}{\epsilon})$
    if $f_i(D) + \nu_i \geq \hat{T}$ **then**
      Output $a_i = \top$.
      Halt.
    else
      Output $a_i = \bot$.
    end if
  end for

aboveThreshold: privacy

Claim: aboveThreshold is $(\varepsilon, 0)$-differentially private.

aboveThreshold: accuracy

Theorem 3.24. For any sequence of k queries $f_1, \ldots, f_k$ such that $|\{i < k : f_i(D) \geq T - \alpha\}| = 0$ (i.e. the only query close to being above threshold is possibly the last one), AboveThreshold$(D, \{f_i\}, T, \epsilon)$ is (α, β) accurate for:

$$\alpha = \frac{8(\log k + \log(2/\beta))}{\epsilon}.$$

to handle c above-threshold queries…

run aboveThreshold c times, restarting after each above-Threshold query

use composition theorems (basic or advanced) to compose privacy guarantees, and union bounding failures of accuracy guarantees

numericSparse

Compose the repeated aboveThreshold algorithm with the Laplace Mechanism to return value of each above-threshold query

numericSparse: privacy

Claim: numericSparse is (ε, δ)-differentially private.

numericSparse: accuracy

Theorem 3.28. For any sequence of k queries $f_1, \ldots, f_k$ such that $L(T) \equiv |\{i : f_i(D) \geq T - \alpha\}| \leq c$, if $\delta > 0$, NumericSparse is (α, β) accurate for:

$$\alpha = \frac{(\ln k + \ln \frac{4c}{\beta})\sqrt{c \ln \frac{2}{\delta}}(\sqrt{512} + 1)}{\epsilon}.$$

If $\delta = 0$, Sparse is (α, β) accurate for:

$$\alpha = \frac{9c(\ln k + \ln(4c/\beta))}{\epsilon}$$