The Sparse Vector Technique

CompSci 590.03
Instructor: Ashwin Machanavajjhala

Announcement

- Project proposal submission deadline is **Fri, Oct 12 noon**.

- How to write the proposal?
  - Just like any paper …
  - … Abstract, Introduction, Notation, Problem Statement, Related Work
  - Instead of algorithms and results sections, you will have section describing how you will solve the problem.

Recap: Laplace Mechanism

Thm: If **sensitivity** of the query is **S**, then adding Laplace noise with parameter **λ** guarantees ε-differential privacy, when

$$\boldsymbol{\lambda = S/\varepsilon}$$

Sensitivity: Smallest number s.t. for any d, d' differing in one entry,

$$|| q(d) - q(d') || \leq S(q)$$

Histogram query: Sensitivity = 2

- Variance / error on each entry = $2\lambda^2 = 2 \times 4/\varepsilon^2$

Cohort Size Estimation Problem

Population of medical patients

Are there at least 200 individuals who are male cancer survivors, between 20-30, who were admitted for surgery

Are there at least 200 male cancer survivors who are between ages of 20 and 30

Are there at least 200 individuals who are male cancer survivors and admitted for surgery

Cohort Size Estimation Problem

- A set of queries {Q1, Q2, Q3, …, Qn}
- Each query Qi : Number of tuples satisfying property pi > τ ?
  - If answer is yes, return the number of tuples satisfying that property And, Researcher performs additional analysis
  - If answer is no, then return NULL.
- Sensitivity of each Qi = 1
- How do we answer using differential privacy?

Duke UNIVERSITY

Cohort Size Estimation Problem

Laplace mechanism:

- Sensitivity of all queries is: n
- For each query: qi’ = Qi(D) + Lap(n/ε)
- Return **qi’** if qi’ > τ
  Return ϕ if qi’ < τ

Accuracy

- We will say that an algorithm is (α, β)-accurate if for a sequence of queries Q1, Q2, …, Qn if with probability > 1-β, the following holds:

$$|q_i' - Q_i(D)| < \alpha \qquad \text{if } q_i' \neq \phi$$

$$Q_i(D) < T + \alpha \qquad \text{if } q_i' = \phi$$

Accuracy of Laplace Mechanism

$$P(|q_i' - Q_i(D)| > \ \alpha) < \beta$$

$$\Rightarrow \int_{\alpha}^{\infty} e^{-\frac{\varepsilon}{n}x}\, dx < \beta$$

$$\Rightarrow \frac{n}{\varepsilon} e^{-\frac{\varepsilon}{n}\alpha} < \ \beta$$

$$\Rightarrow \ \alpha > \ \frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon\beta}\right)$$

Cohort Estimation Problem

- In many exploratory situations, only a small number *c* of the queries actually have a count > τ
- However, accuracy depends on the total number of queries, not just the queries that cross the threshold,
  - Even though we do not return an answer otherwise.
- Is there a mechanism where you need to pay when the count is > τ ?

Duke UNIVERSITY

Sparse Vector Technique

- Set count = 0
- Set τ' = τ + Lap(2/ε)
- For each query: qi' = Qi(D) + Lap(2c/ε)
- If qi' ≥ τ' & count < c,

    count++

    Return **qi'**

  Else if qi' < τ'

    Return ϕ

  Else *// count ≥ c*

    Abort

Use a noisy threshold

Instead of Lap(n/ε)

Answer at most c queries positively

Duke UNIVERSITY

Sparse Vector Technique: Privacy

$$\log\left(\frac{P(M(D) = O)}{P(M(D') = O)}\right)$$

$$= \sum_{i=1}^{n} \log\left(\frac{P(Q_i(D) = o_i | o^{<i})}{P(Q_i(D') = o_i | o^{<i})}\right)$$

$$= \sum_{i:o_i=\emptyset} \log\left(\frac{P(Q_i(D) = \emptyset | o^{<i})}{P(Q_i(D') = \emptyset | o^{<i})}\right)$$

$$+ \sum_{i:o_i\neq\emptyset} \log\left(\frac{P(Q_i(D) = o_i | o^{<i})}{P(Q_i(D') = o_i | o^{<i})}\right)$$

Previous answers
(current answer is not independent of previous answers)

Sparse Vector Technique: Privacy

$$\sum_{i:o_i \neq \emptyset} \log\left(\frac{P(Q_i(D) = o_i | o^{<i})}{P(Q_i(D') = o_i | o^{<i})}\right) \leq \sum_{i:o_i \neq \emptyset} \frac{\varepsilon}{2c} \leq \frac{\varepsilon}{2}$$

At most c queries answered positively

$$want\ to\ show \sum_{i:o_i = \emptyset} \log\left(\frac{P(Q_i(D) = \emptyset | o^{<i})}{P(Q_i(D') = \emptyset | o^{<i})}\right) \leq \frac{\varepsilon}{2}$$

Independent of the number of queries answered with NULL

Sparse Vector Technique: Privacy

- Let $A_Z(D)$ be the set of noise values { vi = qi' – Qi(D) } that result in the observed ϕ answers when τ' = Z.

- If we changed D to D',

$$A_{Z-1}(D') \subseteq A_Z(D) \subseteq A_{Z+1}(D')$$

- If Qi(D) + vi < Z, then Qi(D') + vi ≤ Qi(D) + 1 + vi ≤ Z+1

- If Qi(D') + vi < Z-1, then Qi(D) + vi ≤ Qi(D') + 1 + vi ≤ Z

Sparse Vector Technique: Privacy

- Let $A_Z(D)$ be the set of noise values { vi = qi' – Qi(D) } that result in the observed ϕ answers when τ' = Z.

- If we changed D to D',

$$A_{Z-1}(D') \subseteq A_Z(D) \subseteq A_{Z+1}(D')$$

- Also, from Laplace mechanism,

$$P(\tau' = Z) \leq e^{\frac{\varepsilon}{2}} \cdot P(\tau' = Z + 1)$$

Duke UNIVERSITY

Sparse Vector Technique: Privacy

$$\prod_{i:o_i=\emptyset} P\left(Q_i(D) = \emptyset \middle| o^{<i}\right)$$

$$= \int_{-\infty}^{\infty} P(\tau' = Z) P\left(v_1, v_2, \dots, v_n \in A_Z(D)\right) dZ$$

$$\leq e^{\frac{\varepsilon}{2}} \int_{-\infty}^{\infty} P(\tau' = Z + 1) P\left(v_1, v_2, \dots, v_n \in A_Z(D)\right) dZ$$

$$\leq e^{\frac{\varepsilon}{2}} \int_{-\infty}^{\infty} P(\tau' = Z + 1) P(v_1, v_2, \dots, v_n \in A_{Z+1}(D')) dZ$$

$$= e^{\varepsilon/2} \prod_{i:o_i=\emptyset} P\left(Q_i(D') = \emptyset \middle| o^{<i}\right)$$

Sparse Vector Technique: Privacy

- Pay $c \cdot \varepsilon_1$ (=ε/2) privacy for the questions that have a count greater than the *noisy* threshold.
- You pay ε_2 (=ε/2) privacy for adding noise to the threshold.
- All the questions whose counts are lower than the threshold are answered **for free!**

Sparse Vector Technique: Accuracy

Theorem: For any queries Q1, Q2, …, Qk such that

$$|\{ i : Qi(D) > \tau - \alpha\}| \leq c$$

Then, the sparse vector technique:

1. does not abort, and
2. is (α,β)-accurate for $\alpha = \frac{4c}{\varepsilon}\left(\log n + \log\frac{2}{\beta}\right)$

Recall: Laplace mechanism is (α,β)-accurate for $\alpha > \frac{n}{\varepsilon}\log\left(\frac{n}{\varepsilon\beta}\right)$

Accuracy

- We will say that an algorithm is (α, β)-accurate if for a sequence of queries Q1, Q2, …, Qn if with probability > $1-\beta$ the algorithm does not abort and the following holds:

$$|q_i' - Q_i(D)| < \alpha \qquad \text{if } q_i' \neq \phi, \text{ or } q_i' \geq \tau'$$

$$Q_i(D) < T + \alpha \qquad \text{if } q_i' = \phi, \text{ or } q_i' < \tau'$$

Duke UNIVERSITY

Sparse Vector Technique: Accuracy

- Suppose $\max_i |v_i| + |\tau - \tau'| \leq \alpha$
- When qi' ≠ ϕ, $q'_i - Q_i(D) = v_i \leq \alpha$
- When qi' = ϕ, then qi' < τ'

$$q'_i = Q_i(D) + v_i < \tau' \leq \tau + |\tau - \tau'|$$
$$\Rightarrow Q_i(D) < \tau + |\tau - \tau'| + v_i \leq \tau + \alpha$$

- And, the algorithm always aborts:

$$Q_i(D) < \tau - \alpha \Rightarrow Q_i(D) < \tau - |\tau - \tau'| - |v_i|$$
$$\Rightarrow q'_i = Q_i(D) + v_i < \tau'$$

Sparse Vector Technique: Accuracy

$enough\ to\ prove:$

$$P\left(\max_i |v_i| + |\tau - \tau'| > \alpha\right) < \beta$$

$$If\ Y \sim Lap(b),\ \ then\ P(Y > t \cdot b) < e^{-t}$$

$$P\left(|\tau - \tau'| > \frac{\alpha}{2}\right) < e^{-\frac{\varepsilon\alpha}{4}} = \frac{\beta}{2}$$

$$P\left(\max_i |v_i| > \frac{\alpha}{2}\right) < k \cdot e^{-\frac{\varepsilon\alpha}{4c}} = \frac{\beta}{2}$$

Summary of Sparse Vector Technique

- If you have many low sensitivity queries, and you only expect a few of the queries to be useful.

- Sparse vector techniques allows you to pay only for the positively answered queries.

- Much smaller error than the Laplace mechanism.

Next Class

- Multiplicative Weights Algorithms
  - General paradigm for algorithm design

  - Application to privately answering queries
  - Application to privately publishing a dataset