Post-processing outputs for better utility

CompSci 590.03
Instructor: Ashwin Machanavajjhala
Announcement

• Project proposal submission deadline is Fri, Oct 12 noon.
Recap: Differential Privacy

For every pair of inputs that differ in one value:

\[ D_1 \quad D_2 \]

Adversary should not be able to distinguish between any \( D_1 \) and \( D_2 \) based on any \( O \):

\[
\log\left( \frac{\Pr[A(D_1) = O]}{\Pr[A(D_2) = O]} \right) < \varepsilon \quad (\varepsilon > 0)
\]

For every output ...

\[ O \]
Recap: Laplacian Distribution

Privacy depends on the \( \lambda \) parameter

\[ h(\eta) \alpha \exp(-\eta / \lambda) \]

Mean: 0, Variance: 2 \( \lambda^2 \)

Database -> True answer \( q(d) \) -> \( q(d) + \eta \) -> Researcher

Laplace Distribution – \( \text{Lap}(\lambda) \)

Lecture 10: 590.03 Fall 12
Recap: Laplace Mechanism

**Thm:** If sensitivity of the query is $S$, then the following guarantees $\varepsilon$-differential privacy.

$$\lambda = S/\varepsilon$$

**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 = $2 \times 4/\varepsilon^2 = O(1/\varepsilon^2)$
This class

• What is the optimal method to answer a batch of queries?
How to answer a batch of queries?

• Database of values \{x_1, x_2, \ldots, x_k\}

• Query Set:
  – Value of \(x_1\) \(\eta_1 = x_1 + \delta_1\)
  – Value of \(x_2\) \(\eta_2 = x_2 + \delta_2\)
  – Value of \(x_1 + x_2\) \(\eta_3 = x_1 + x_2 + \delta_3\)

• But we know that \(\eta_1\) and \(\eta_2\) should sum up to \(\eta_3\)!
Two Approaches

• **Constrained inference**
  – Ensure that the returned answers are consistent with each other.

• **Query Strategy**
  – Answer a different set of *strategy* queries A
  – Answer original queries using A

  – Universal Histograms
  – Wavelet Mechanism
  – Matrix Mechanism
Two Approaches

• **Constrained inference**
  – Ensure that the returned answers are consistent with each other.

• **Query Strategy**
  – Answer a different set of *strategy* queries A
  – Answer original queries using A
    – Universal Histograms
    – Wavelet Mechanism
    – Matrix Mechanism
Constrained Inference

DATA OWNER

Private Data

\[ Q(I) = q \]

ANALYST

\[ \tilde{q} \]

Constrained Inference

\[ \overline{q} \]
Constrained Inference

• Let $x_1$ and $x_2$ be the original values. We observe noisy values $\eta_1$, $\eta_2$ and $\eta_3$

• We would like to reconstruct the best estimators $y_1$ (for $x_1$) and $y_2$ (for $x_2$) from the noisy values.

• That is, we want to find the values of $y_1$, $y_2$ such that:

$$\min (y_1-\eta_1)^2 + (y_2 - \eta_2)^2 + (y_3 - \eta_3)^2$$

s.t., $y_1 + y_2 = y_3$
**Definition 2.4 (Minimum $L_2$ solution).** Let $Q$ be a query sequence with constraints $\gamma_Q$. Given a noisy query sequence $\tilde{q} = \tilde{Q}(I)$, a minimum $L_2$ solution, denoted $\tilde{q}$, is a vector $\tilde{q}$ that satisfies the constraints $\gamma_Q$ and at the same time minimizes $\|\tilde{q} - q\|_2$. 
Sorted Unattributed Histograms

• Counts of diseases
  – (without associating a particular count to the corresponding disease)

• Degree sequence: List of node degrees
  – (without associating a degree to a particular node)

• Constraint: The values are sorted
Sorted Unattributed Histograms

True Values: 20, 10, 8, 8, 8, 5, 3, 2
Noisy Values: 25, 9, 13, 7, 10, 6, 3, 1 (noise from \text{Lap}(1/\varepsilon))

\[
\begin{align*}
\text{minimize} & \quad ||\tilde{s} - \bar{s}||_2 \\
\text{s.t.} & \quad \forall i, \tilde{s}_i \leq \bar{s}_{i-1}
\end{align*}
\]

\[
\bar{s}_k = \min_{j \in [k,n]} \max_{i \in [1,j]} \frac{(\tilde{s}_i + \tilde{s}_{i+1} + \ldots + \tilde{s}_j)}{j - i + 1}
\]

Proof?:

Lecture 10 : 590.03 Fall 12
Sorted Unattributed Histograms
Sorted Unattributed Histograms

- $n$: number of values in the histogram
- $d$: number of distinct values in the histogram
- $n_i$: number of times $i^{th}$ distinct value appears in the histogram.

**Theorem 2.** There exist constants $c_1$ and $c_2$ independent of $n$ and $d$ such that

$$\text{error}(\overline{S}) \leq \sum_{i=1}^{d} \frac{c_1 \log^3 n_i + c_2}{\epsilon^2}$$

Thus $\text{error}(\overline{S}) = O(d \log^3 n/\epsilon^2)$ whereas $\text{error}(\tilde{S}) = \Theta(n/\epsilon^2)$. 
Two Approaches

• **Constrained inference**
  – Ensure that the returned answers are consistent with each other.

• **Query Strategy**
  – Answer a different set of *strategy* queries A
  – Answer original queries using A
  
  – **Universal Histograms**
  – **Wavelet Mechanism**
  – **Matrix Mechanism**
Query Strategy

Private Data

Original Query Workload

A

Strategy Query Workload

A(I)

Differential Privacy

Noisy Strategy Answers

Â(I)

Noisy Workload Answers

W

Ŵ(I)
Range Queries

- Given a set of values \{x_1, x_2, \ldots, x_n\}
- Range query: \(q(j,k) = x_j + \ldots + x_k\)

Q: Suppose we want to answer all range queries?

Strategy 1: Answer all range queries using Laplace mechanism

- \(O(n^2/\varepsilon^2)\) total error.
- May reduce using constrained optimization ...
Range Queries

- Given a set of values \{x_1, x_2, \ldots, x_n\}
- Range query: \(q(j,k) = x_j + \ldots + x_k\)

Q: Suppose we want to answer all range queries?

Strategy 1: Answer all range queries using Laplace mechanism

- Sensitivity = \(O(n^2)\)
- \(O(n^4/\varepsilon^2)\) total error across all range queries.
- May reduce using constrained optimization ...
Range Queries

• Given a set of values \{x_1, x_2, \ldots, x_n\}
• Range query: \( q(j,k) = x_j + \ldots + x_k \)

Q: Suppose we want to answer all range queries?

Strategy 2: Answer all xi queries using Laplace mechanism
Answer range queries using noisy xi values.

• \( O(1/\epsilon^2) \) error for each xi.
• \( \text{Error}(q(1,n)) = O(n/\epsilon^2) \)
• Total error on all range queries: \( O(n^3/\epsilon^2) \)
Universal Histograms for Range Queries

Strategy 3:

Answer *sufficient statistics* using Laplace mechanism
Answer range queries using noisy sufficient statistics.

[Hay et al VLDB 2010]
Universal Histograms for Range Queries

• Sensitivity: $\log n$
• $q(2,6) = x_2 + x_3 + x_4 + x_5 + x_6$
  $= x_2 + x_{34} + x_{56}$

Error = $2 \times 5\log^2 n/\epsilon^2$
Error = $2 \times 3\log^2 n/\epsilon^2$
Universal Histograms for Range Queries

- Every range query can be answered by summing at most $\log n$ different noisy answers.
- Maximum error on any range query = $O(\log^3 n / \varepsilon^2)$
- Total error on all range queries = $O(n^2 \log^3 n / \varepsilon^2)$
Universal Histograms & Constrained Inference

- Can further reduce the error by enforcing constraints
  \[ x_{1234} = x_{12} + x_{34} = x_1 + x_2 + x_3 + x_4 \]

\[
\begin{align*}
\text{minimize} & \quad \sum_{\nu} (\tilde{c}(\nu) - \bar{c}(\nu))^2 \\
\text{s.t.} & \quad \bar{c}(\nu) = \sum_{u \in \text{child}(\nu)} \bar{c}(u)
\end{align*}
\]

- 2-pass algorithm to compute a consistent version of the counts
Universal Histograms & Constrained Inference

• Pass 1: (Bottom Up)

\[ z(v) = \begin{cases} 
\bar{c}(v), & \text{if leaf node} \\
\alpha \cdot c(v) + (1 - \alpha) \sum_{u=\text{child}(v)} z(u) 
\end{cases} \]

• Pass 2: (Top down)

\[ \bar{c}(v) = \begin{cases} 
z(v), & \text{if root node} \\
z(v) + \frac{1}{2} \left( \bar{c}(v) - \sum_{u=\text{child}(v)} z(u) \right) 
\end{cases} \]
Universal Histograms & Constrained Inference

• Resulting consistent counts
  – Have lower error than noisy counts (upto 10 times smaller in some cases)
  – Unbiased estimators
  – Have the least error amongst all unbiased estimators
Next Class

• Constrained inference
  – Ensure that the returned answers are consistent with each other.

• Query Strategy
  – Answer a different set of strategy queries A
  – Answer original queries using A
    – Universal Histograms
    – Wavelet Mechanism
    – Matrix Mechanism