Lecture 13: Multiplicative Weights Algorithm for Synthetic Data

CompSci 590.03 *Ashwin Machanavajjhala*

1 Introduction

In this lecture, we will first present a multiplicative weights algorithm for generating synthetic data without ensuring privacy (Section 2). We will modify this algorithm to ensure that the synthetic data that is generated satisfies ϵ-differential privacy (Section 3). The private algorithm originally appeared in Hardt et al, "A simple and practical algorithm for differentially private data release", *NIPS* 2012.

2 Multiplicative Weights & Synthetic Data

Input:

- D, a database of of n records, where each record drawn from a domain $\mathcal{D}$.
- Q, a set of linear queries on D. That is, each query q can be written as $q(D) = \sum_{x \in D} f_q(x)$, and $f_q : \mathcal{D} \rightarrow [-1, +1]$. Henceforth, we will abuse the notation $q(x)$ to also denote function value $f_q(x)$.

Output:

- $A : \mathcal{D} \rightarrow R$, a function that assigns weights to each element in the domain such that:
  - $\sum_{x \in \mathcal{D}} A(x) = n$, and
  - for each query $q \in Q$, $|q(A) - q(D)| \leq \alpha$, where $q(A) = \sum_{x \in \mathcal{D}} q(x) \cdot A(x)$.

Solution:

1. *Initialization:* A_0 is a function that assigns $n/\mathcal{D}$ to every element $x \in \mathcal{D}$.
2. *Iteration i:*
   (a) Pick the query $q_i = argmax_{q \in Q} |q(D) - q(A_{i-1})|$
   (b) Compute $m = q_i(D)$
   (c) *Update Weights:*

$$A_i(x) \propto A_{i-1}(x) \cdot exp(q_i(x) \cdot (m - q_i(A_i))/2n)$$

3. Stop after T iterations. Return $\frac{1}{T} \sum_{i=1}^{T} A_{i-1}$

Error Analysis:

Claim 1.

$$\max_{q \in Q} |q(A) - q(D)| \leq 2n\sqrt{\frac{\log |\mathcal{D}|}{T}} \tag{1}$$

PROOF:

$$\begin{aligned}
\max_{q \in Q} |q(A) - q(D)| &\leq \max_{q \in Q} |q(\frac{1}{T}\sum_{i=1}^{T} A_{i-1}) - q(D)| \\
&\leq \frac{1}{T}\sum_{i=1}^{T} \max_{q \in Q} |q(A_{i-1}) - q(D)| \\
&\leq \frac{1}{T}\sum_{i=1}^{T} |q_i(A_{i-1} - q_i(D)|
\end{aligned}$$

Consider the following potential function:

$$\Phi(i) \;=\; \sum_{x \in \mathcal{D}} \frac{D(x)}{n} \log \frac{D(x)}{A_i(x)}$$

We can easily show that $\Phi(i) \geq 0$ and $\Phi(0) \leq \log |\mathcal{D}|$. We also show in Claim 2 that

$$\forall i,\, \Phi(i-1) - \Phi(i) \;\geq\; \left(\frac{q_i(A_{i-1}) - q_i(D)}{2n}\right)^2$$

Therefore, we have:

$$\begin{aligned}
\max_{q \in Q} |q(A) - q(D)| &\leq \frac{1}{T}\sum_{i=1}^{T} |q_i(A_{i-1} - q(D)| \\
&\leq \frac{1}{T}\sum_{i=1}^{T} 2n\sqrt{\Phi(i-1) - \Phi(i)} \\
&\leq 2n\sqrt{\frac{1}{T}\sum_{i=1}^{T}(\Phi(i-1) - \Phi(i))} \quad \text{cauchy-schwartz inequality} \\
&\leq 2n\sqrt{\frac{\log |\mathcal{D}|}{T}}
\end{aligned}$$

□

Claim 2.

$$\forall i,\, \Phi(i-1) - \Phi(i) \;\geq\; \left(\frac{q_i(A_{i-1}) - q_i(D)}{2n}\right)^2 \tag{2}$$

PROOF:

$$\begin{aligned}
\Phi(i-1) - \Phi(i) &= \sum_{x \in \mathcal{D}} \frac{D(x)}{n} \log \frac{A_i(x)}{A_{i-1}(x)} \\
\frac{A_i(x)}{A_{i-1}(x)} &= \frac{1}{z_i} e^{q(x) \cdot (q(D) - q(A_{i-1}))/2n} \\
&= \frac{1}{z_i} e^{q(x) \cdot \eta_i} \\
\Phi(i-1) - \Phi(i) &= \sum_{x \in \mathcal{D}} \frac{D(x)}{n} \left(q_i(x) \eta_i - \log z_i \right) \\
&= \frac{q_i(D)}{n} \eta_i - \log z_i
\end{aligned}$$

where, z_i is the normalization factor, and $\eta_i = (q(D) - q(A_{i-1}))/2n$.

$$\begin{aligned}
z_i &= \sum_{x \in \mathcal{D}} e^{q_i(x)\eta_i} \frac{A_{i-1}(x)}{n} \\
&\leq \sum_{x \in \mathcal{D}} (1 + q_i(x)\eta_i + q_i(x)^2 \eta_i^2) \frac{A_{i-1}(x)}{n} \quad \text{since } e^x \leq (1 + x + x^2) \\
&\leq \sum_{x \in \mathcal{D}} (1 + q_i(x)\eta_i + \eta_i^2) \frac{A_{i-1}(x)}{n} \quad \text{since } q_i(x)^2 \leq 1 \\
&= 1 + \eta_i^2 + \frac{\eta_i q_i(A_{i-1})}{n} \\
\log z_i &\leq z_i - 1 \quad \text{since } \log(1+x) \leq x \\
&\leq \eta_i^2 + \frac{\eta_i q_i(A_{i-1})}{n}
\end{aligned}$$

Therefore, we get:

$$\begin{aligned}
\Phi(i-1) - \Phi(i) &= \frac{q_i(D)}{n} \eta_i - \log z_i \\
&\geq \frac{q_i(D)}{n} \eta_i - \frac{q_i(A_{i-1})}{n} \eta_i - \eta_i^2 \\
&= \frac{(q_i(D) - q_i(A_{i-1}))^2}{2n^2} - \frac{(q_i(D) - q_i(A_{i-1}))^2}{4n^2} \\
\Phi(i-1) - \Phi(i) &\geq \left(\frac{q_i(D) - q_i(A_{i-1})}{2n} \right)^2
\end{aligned}$$

□

3 Private Synthetic Data Generation using Multiplicative Weights

In order to guarantee ϵ-differential privacy, we need to modify the algorithm so that query answers returned from the true database D are done in a privacy preserving manner. The true database is consulted twice in each iteration in the algorithm:

- To compute query q_i with the maximum error on the current approximation A_{i-1}. This can be made private by choosing the query using the Exponential Mechanism.
- To compute the answer $q_i(D)$. We can compute a noisy answer using the Laplace Mechanism.

We present private algorithm below. The changes to the algorithm from the previous section are *italicized*.

Solution:

1. *Initialization:* A_0 is a function that assigns $n/\mathcal{D}$ to every element $x \in \mathcal{D}$.
2. *Iteration i:*
   (a) Pick the query q_i *using Exponential Mechanism with parameter* $\epsilon/2T$*, and a score function* $|q(A_{i-1}) - q(D)|$.
   (b) Compute $m = q_i(D)$ *using the Laplace Mechanism with parameter* $\epsilon/2T$.
   (c) *Update Weights:*

$$A_i(x) \propto A_{i-1}(x) \cdot exp(q_i(x) \cdot (m - q_i(A_i))/2n)$$

3. Stop after T iterations. Return $\frac{1}{T}\sum_{i=1}^{T} A_{i-1}$

Privacy Analysis: The above algorithm satisfies ϵ-differential privacy. This follows from the composability of differential privacy – each iteration expends $2 \cdot \epsilon/2T = \epsilon/T$ of the privacy budget, and there are at most T iterations.

Error Analysis: We can guarantee almost the same error bound with high probability plus an additive factor due the error induced by the Exponential and Laplace mechanisms. The analysis is very similar to Claim 1. We refer the reader to Hardt et al for the complete proof.

Claim 3. *With probability at least* $1 - 2T/|Q|$,

$$\max_{q \in Q} |q(A) - q(D)| \leq 2n\sqrt{\frac{\log |\mathcal{D}|}{T}} + \frac{10T \log |Q|}{\epsilon} \tag{3}$$

PROOF: We can guarantee almost the same error bound with high probability plus an additive factor due the error induced by the Exponential and Laplace mechanisms.

$$\begin{aligned} \max_{q \in Q} |q(A) - q(D)| &\leq \max_{q \in Q} |q(\frac{1}{T}\sum_{i=1}^{T} A_{i-1}) - q(D)| \\ &\leq \frac{1}{T}\sum_{i=1}^{T} \max_{q \in Q} |q(A_{i-1}) - q(D)| \end{aligned}$$

Unlike in the previous analysis, we are not guaranteed that the q_i picked in each iteration has the maximum error (call it $maxerr_i$). However, we know that when using Exponential mechanism,

$$P(|q_i(A_{i-1} - q_i(D)| < maxerr_i - r) \quad < \quad |Q|e^{-fracer4T}$$

When, $r = 8T \log |Q|/\epsilon$, we get that with probability at least $1 - 1/|Q|$,

$$maxerr_i \ \leq \ |q_i(A_{i-1} - q_i(D)| + 8T \log |Q|/\epsilon \tag{4}$$

Similarly, we add noise $v \sim Lap(2t/\epsilon)$ when computing $q(D)$. It is easy to show that

$$P(v > 2T \log |Q|/\epsilon) < e^{-\log |Q|} < 1/|Q| \tag{5}$$

Putting together Equations 4 and 5, and the analysis from the previous section, we get the required result. We refer the reader to Hardt et al for the complete proof. □