Modern Topics in Statistical Learning Theory Spring 2024

Lecture 12 Reconstruction Attack and Differential Privacy

Prof. Qi Lei *Scribe: Gail Batutis, Jiyang Du*

1 Motivation

There are many different kinds of privacy, but currently differential privacy is the defining criteria, and almost ubiquitous. Consider a typical scenario involving a hospital that reports certain disease statistics, such as the number of AIDS cases relative to the patient count. The common belief is that this statistical information can be safely published without compromising individual privacy. However, an example of how this can go wrong is when statistics reveal that 4 out of 2436 patients have AIDS one week, and the following week the number changes to 5 out of 2437. The inference that the newly added individual has AIDS illustrates a privacy breach. Differential privacy addresses this issue by ensuring that the difference in observed data, such as statistical reports, does not reveal specific private information about individuals.

Another practical example involves interactions with AI models like ChatGPT. Suppose you contribute to the training data of ChatGPT 3.5 and then interact with the upgraded ChatGPT 4. The updates in the model, driven by your input, might lead to concerns about whether the model is memorizing specific training data. Interacting with different versions of the model could potentially expose parts of the training data, highlighting the need for robust differential privacy measures to prevent such unintended disclosures.

2 Reconstruction Attack

Why do we introduce attack before defining data privacy?

1. the idea of an attack can exist without proper definition of privacy
2. it can help us evaluate whether a notion of privacy is too strong or too weak:

    - notion of privacy (with Definition A) is satisfied but still allows successful attacks ⇒ Def A is too weak.
    - notion of privacy with Def B is not satisfied but no attacks can succeed ⇒ Def B is excessively stringent.

Understanding the various types of attacks and the scenarios in which they occur is essential to effectively define and uphold privacy standards.

Example. We look at a reconstruction attack on synthetic US Census data[3]. In this as in many cases, sharing statistical information can be a public benefit. Statistical information does not need to be private. However, personal information needs to be protected, and statistical information may still reveal personal information. See Table 1 for ”fictional statistical data for a fictional block” [3]. When count is ≤ 2, that data entry (row) is hidden, since we can recover all ages for that group using mean and average.

Table 1: Synthetic Data from the US census

Statistical Group	**Description**	**Count**	**Age Median**	**Age Mean**
1A	Total Population	7	30	38
2A	Female	4	30	33.5
2B	Male	3	30	44
2C	Black/AA	4	51	48.5
4A	Black/AA Female	3	36	36.7

Note that from Groups 2C and 4A, we can still infer that the Black/AA male is of age 84 = 48.5*4-36.7*3, and therefore the data is not private[3].

A general attack that tries to recover all microdata from published statistics works as follows:

1. treat the attributes of each person as a collection of variables
2. formulate each row in the published statistical table as mathematical constraints/equations
3. find a feasible solution that satisfies all constraints (we can solve the problem using integer programming)

If the statistics are highly constraining, the solution is unique, and we can recover information on every person.

Example. We now examine the Dinur-Nissim database reconstruction attack[1].
In the database shown in Table 2, each row is a data point and each column is a feature. The first four columns are the identifier information and the last column is the sensitive information.

Table 2: DN Database Reconstruction Attack

Name	**Postal Code**	**Date of Birth**	**Sex**	**Has Disease?**
Alice	K8V7R6	5/2/1984	F	1
Bob	V5K5J9	2/8/2001	M	0
Charlie	V1C7J	10/10/1954	M	1
David	R4K5TI	4/4/1944	M	0
Eve	G7N8Y3	1/1/1980	F	1

In our threat model, we ask: how much capacity does the adversary have? i.e., how hard it is to get the secret bit from the statistics of the queried row?

Let $d =$(Has disease?), $A(S) = \sum_i d_i =$ (# of disease = 1 in set S). For a set S, the data curator will generate the answer r(S). e.g $S = \{$ Alice, Bob, Charlie$\}$, r(S) = 2.

- If r(S) is independent of A(S), we wouldn't reveal any information, but this is in some cases impractical, the utility of the database will be ruined.
- If r(S) = A(S), then we can reconstruct d easily. Let n be the number of rows in the table. Then
$$d = \begin{bmatrix} r(\{1\}) \\ r(\{2\}) \\ \vdots \\ r(\{n\}) \end{bmatrix}$$
- If $r(S) + A(S) +$ bounded noise, i.e. $|r(S) - A(S)| \leq E$ for some constant E, this is a more interesting case.

From [1], we have the following result:

Theorem 1. *Let $n = \#$ of entries. If the adversary is allowed to ask 2^n subset queries, and the curator adds noise with bound E, then the adversary can reconstruct all but $4E$ entries.*

Remark:

- Even when E scales with n, recovery ratio can be high. Let $E = n/401$, then the bound tell us we can reconstruct $n - \frac{n}{100}$ rows, i.e 99% of the data!
- 2^n is not a tight upper bound for the number of queries needed.

Proof. Attack as follows:

Algorithm 1 Attack

```
ask all $2^n$ questions
for $c \in \{0,1\}^n$ do
    if $\exists S$ st $|\sum_{i \in S} c_i - r(S)| > E$ then
        rule out c
    else
        output c                    ▷ This is equivalent to output any c that's not ruled out
```

Easy to see that the true answer d is not ruled out. Let $I_0 = \{i : d_i = 0\}, I_1 = \{i : d_i = 1\}$. Note the output is not necessarily unique. For any c in the output, we have

$$|\sum_{i \in I_0} c_i - r(I_0)| \leq E, \quad |r(I_0) - \sum_{i \in I_0} d_i| \leq E$$

thus c, d differs at most $2E$ on I_0. By same argument, we have c, d differs at most $2E$ on I_1. □

Theorem 2. *[1] If the adversary is allowed to ask $O(n)$ subset queries and $E = O(\alpha\sqrt{n})$, then one can reconstruct all but $O(\alpha^2)$ entries.*

Remark:

- Previously for $E = n/401 = O(n) \propto \alpha\sqrt{n}$, $\alpha = O(\sqrt{n})$. Thm 2 gives the bound O(n), which is consistent with the result from Thm 1, $n/100$.
- Thm 2 is tighter in terms of the number of subset queries allowed.
- In the next lecture, we will see that with privacy guarantees, adding noise of $O(\sqrt{n})$ is sufficient. This bound on E is tight and the best possible (cannot get the same result with a smaller E).

Proof not presented here.

3 Differential Privacy

Differential privacy is a concept centered around a mathematical framework where an algorithm, M, maps a set of observations, $\mathcal{X}^n$, to an output, $\mathcal{Y}$. The output, $\mathcal{Y}$, could represent various statistical data like the summation of people with a specific disease or a model trained from the data. The goal of (ϵ-pure) differential privacy is to ensure that for all pairs of neighboring datasets X and X' and T $\subseteq \mathcal{Y}$, the probability of $M(X) \in T$ is not significantly different from $M(X') \in T$.

Definition 3. *[2]Let $\mathcal{X}^n$ be a collection of datasets. For an algorithm $M : \mathcal{X}^n \to \mathcal{Y}$, we say M provides ϵ-(pure) differential privacy if for all "neighboring" $X, X' \in \mathcal{X}^n$ (X, X' differ on exactly one entry), and any $T \subseteq \mathcal{Y}$,*

$$\mathbb{P}(M(X) \in T) \leq e^{\epsilon}\, \mathbb{P}(M(X') \in T).$$

Note

- the randomness comes from the algorithm M
- switching role of X, X', from the definition we have

  $$e^{\epsilon}\, \mathbb{P}(M(X') \in T) \leq e^{2\epsilon}\, \mathbb{P}(M(X) \in T)$$

  This implies $1 \leq e^{2\epsilon}$, so ϵ needs to be nonnegative.
- this is a worst-case guarantee as we have a uniform bound for all neighboring X, X' and any T
- as ϵ decreases, the inequality above gives a tighter bound, resulting in stronger DP. Practically, ϵ is typically within the range of 0.5 to 5; values outside this range may render the privacy either ineffective or suspiciously stringent.
- we write the multiplicative as e^{ϵ} for convenience. In the more general case where we have more than one secret bit, we could have a different bound for each bit, and instead of the result being a product of multiple epsilon values, the multiplicative would become $\exp(\sum_i \epsilon_i)$

Furthermore, differential privacy is symmetric, meaning that the datasets X and X' can be interchanged without affecting the privacy guarantees. In practice, this manifests as an observer being unable to distinguish between two datasets based on the observed outputs, such as the number of disease cases, thereby maintaining privacy across all neighboring datasets.

Differential Privacy Guarantee:
The differential privacy (DP) guarantee fundamentally complicates the exact reconstruction of a dataset.

Consider the following hypothesis testing problem.

$$H_0 : \text{the underlying dataset is } X$$
$$H_1 : \text{the underlying dataset is } X'$$

Then from the definition of ϵ-DP we have a bound for the test statistics in the likelihood ratio test

$$\frac{P[\text{observation} = Y|H_0]}{P[\text{observation} = Y|H_1]} \in (e^{-\epsilon}, e^{\epsilon})$$

This statistical framework makes it challenging to decisively refute any hypothesis, rendering any attempts to do better than a random guess unlikely. In essence, if reconstructing data from observations is more complex than hypothesis testing, the DP guarantee asserts that one cannot perform hypothesis testing effectively, thereby safeguarding against the possibility of reconstruction.

In the next lecture, we will see a specific choice of algorithm M and that DP is too strong for modern reconstruction problems.

4 Acknowledgement

Part of the content was from Prof. Gantam Kamath's notes for CS860.

References

[1] I. Dinur and K. Nissim. Revealing information while preserving privacy. pp. 202–210, 06 2003. doi: 10.1145/773153.773173

[2] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9, 01 2013. doi: 10.1561/0400000042

[3] S. Garfinkel, J. Abowd, and C. Martindale. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62:46–53, 02 2019. doi: 10.1145/3287287