Privacy I

Lecturer: Kamalika Chaudhuri

April 24, 2020

1 Why do we need Rigorous Privacy?

A lot of machine learning is currently done on sensitive data. Much of AI for healthcare involves processing, analyzing and learning from large volumes of medical data – which is considered very sensitive. Bioinformatics is done on human genomic data – which again could be highly sensitive as it reveals information about ourselves. Finally, user behavior data – financial transactions, search logs, and other behavioural data are highly sensitive as well. In the next lectures, we will talk about the different notions of privacy, and how to preserve them while training machine learning models on sensitive data.

1.1 Anonymization

The simplest kind of privacy mechanism that one can think of is anonymization – remove names and addresses, and other *obviously* identifying bits, and release the data. Unfortunately this does not work as well as planned. A well-known example is the 2006 AOL fiasco – where AOL publicly released a large portion of its search logs with the intent of helping researchers develop better search algorithms. They replaced user names with random numbers, and for each user, now represented by a random number, they published a list of search queries. The next day, there was an article in the New York Times with interviews from some of the people in dataset – it took a reporter just a few hours to identify a handful of people in the data! After this, AOL retracted the dataset. This incident is just one example, and the privacy literature is replete with many such cases of broken anonymization.

Why does anonymization not work? The reason is that even moderately dimensional data about people tends to be very unique, and these unique combination of features act as *pseudo-identifiers* that can be combined with linkage information to precisely deidentify people. As a concrete example, suppose UCSD decides to publish an anonymized table of salaries of its employees, and you see the following row:

It is very clear that there is only one person that fits this description, and that is me. Not just that, the linkage information – that links this row to me is on our department's website, which has a list of faculty profiles with photos.

Position	Department	Gender	Ethnicity	Salary
Faculty	CSE	Female	SE Asian	-

1.2 Releasing Raw Statistics

Not just releasing data, releasing a lot of raw statistics based on small amounts of data can also lead to privacy leakage. Consider as an example a study where there are two groups of people – cancer patients and healthy people, and we gather genetic data from each, and suppose we compute and publish two tables of cross-correlations of SNP values, one for each group. If there are just a few people in each group, and many more SNPs than people, then one can reverse engineer the tables and find out what the genetic profiles are. An adversary who has access to these tables (which are raw statistics) and knows a few of Alice's SNPs can determine if Alice is in either of these tables, and if yes, whether she is in the cancer table or the healthy table.

Thus, even if we are releasing statistics and not raw data, it is unsafe to release too many statistics that are constructed out of small datasets. To measure what is safe, and what constitutes small data, we need a rigorous notion of privacy, which we look at next.

2 Differential Privacy

We next look at differential privacy – which is considered now the gold standard in privacy-preserving data analysis. The main idea behind differential privacy is that we would like aggregate information about a population to be computable, while hiding individual information about people. For example, we may want to discover population-level patterns, such as smoking causes cancer, while hiding the fact that a specific person who participated in the study smokes. We also want the privacy definition to be robust to side information that an adversary might have.

2.1 The Central Privacy Setting

Before we get to the precise definition, let us first clarify the setting. The standard differential privacy definition is in what is called the central privacy setting, which will be the focus of most of this class. As an aside, there is also a local privacy setting, which we might talk about briefly.

In the central setting, users provide their raw data directly to a trusted curator, who aggregates the data, and releases a sanitized aggregate to the public. The privacy barrier is thus between the curator and the public; an adversary has no insight into what goes on inside the curator, and has only access to its output. An example of this setting is the US Census which is planning to incorporate differential privacy in 2020. Residents still send their raw data directly to the census by filling out the forms; the census statisticians then convert this data into sanitized tables which preserve privacy.

2.2 The Definition

The main idea behind differential privacy is that the participation of a single person in a dataset should not make a difference to the probability of any outcome. For example, whether you decide to fill out the census form or not should not make a difference to the probability of anything that might happen to you. Thus regardless of what prior information an adversary may have, the probability of any outcome remains the same whether Alice participates in a dataset or not. This incentivizes Alice to participate – as she has not much to lose by not doing so.

How is this kind of guarantee provided? In differential privacy, it is through randomness. Any differentially private algorithm is a randomized algorithm, whose randomness might come from different sources such as explicit noise addition and subsampling. If A is a differentially private algorithm, then the output of A is drawn from a distribution. The goal of differential privacy is to make sure that if we take a dataset D, and if we switch out Alice and replace her with Bob to get a second dataset D', then the output distributions for $A(D)$ and $A(D')$ are close. Here, closeness is measured by pointwise likelihood ratio, which is required to be bounded.

More formally, the definition is as follows.

Definition 1 (Differential Privacy)**.** *A (randomized) algorithm A whose output lies in a set $\mathcal{S}$ is ϵ-differentially private if for all datasets D and D' that differ in the private value of a single person, and for all $S \subseteq \mathcal{S}$, we have:*

$$\Pr(A(D) \in S) \leq e^{\epsilon} \Pr(A(D') \in S)$$

ϵ is a privacy parameter where low ϵ means high privacy; ϵ is sometimes called the privacy budget.

Observe that this kind of participation-based definition is rather subtle, and has some interesting implications. Consider, for example, an adversary who is Alice's health insurance provider and who would like to find out if Alice has cancer or not so that they can raise her premiums. Being her health insurance provider, this adversary has some prior knowledge – suppose they know her genetic profile and that she smokes. Consider two cases. In the first case, Alice has participated in the study described in Section 1.2, and the adversary finds out that Alice is in the cancer group by reverse-engineering the tables. This is a breach according to differential privacy – if Alice did not participate in the study, then no harm would come to her.

Now suppose in the second case, a study comes out from a different country (from where Alice lives) which shows that smoking causes cancer. The adversary reads this study and concludes that Alice might soon develop cancer and decides to raise her premiums. In both cases, harm comes to Alice, but unlike the first, the second study does not violate Alice's privacy according to differential privacy terms. If the findings of the study is stable, then whether Alice participates in it or not makes no difference; even if Alice did not participate in it, her health insurance provider would choose to raise her premiums.

A slightly relaxed notion is approximate differential privacy.

Definition 2 (Approximate Differential Privacy)**.** *A (randomized) algorithm A whose output lies in a set $\mathcal{S}$ is said to satisfy (ϵ, δ)-approximate differential privacy if for all datasets D and D' that differ in a single person's private value, and all $S \subseteq \mathcal{S}$, we have:*

$$\Pr(A(D) \in S) \leq e^{\epsilon} \Pr(A(D') \in S) + \delta$$

Approximate differential privacy adds an extra parameter δ; again small δ implies better privacy.

2.3 Properties

Differential privacy has several very nice properties that makes it highly desirable for a number of applications; not all privacy definitions have these properties.

- **Resistance to Side Information.** Because of the way privacy is defined, the notion is resistant to side information. For any side information an adversary may have, participation of a single person will not make a difference. This could be important in applications such as the census where the adversary could be arbitrarily powerful.

- **Post-processing Invariance.** Post-processing invariance means that the privacy risk to an individual in the data does not increase if we post-process the output of an ϵ-differentially private algorithm in any way provided that we do not go back to the original sensitive data. This property is good to have in applications such as the census. Once the tables are released, the census has no control over what various government agencies or the public will do with them.

- **Graceful Composition.** Privacy composition refers to the increase in privacy risk as multiple private releases are made based on the same sensitive dataset. Differential privacy obeys graceful composition, in the sense that the privacy risk increases slowly with multiple releases. The simplest form of composition states that the increase in linear – if two releases are (ϵ_1, δ_1) and (ϵ_2, δ_2) differentially private, then their union is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private. We will look at slightly better bounds later on in the class.

- **Group Privacy.** The differential privacy definition addresses the effect of participation of single people; however in differential privacy, the effect of participation of groups is also bounded. Specifically, if D and D' differ in the private values of k people, then we have:

  $$\Pr(A(D) \in S) \leq e^{k\epsilon} \Pr(A(D') \in S) + k\delta$$

  So the parameters go up by a factor of k.

2.4 Some Basic Mechanisms

There is a large body of work in designing mechanisms that satisfy differential privacy with increasingly better utility. We next discuss two basic ones – the Global Sensitivity Mechanism, and the Exponential Mechanism.

The problem setting is as follows. There is a private dataset D, and we would like to approximately compute a function $f(D)$ on it with ϵ-differential privacy. For simplicity, we look at scalar functions f; similar mechanisms apply to vector functions but need more complicated calculations.

2.4.1 The Global Sensitivity Mechanism

In the Global Sensitivity Mechanism, we add a noise term Z to $f(D)$ to preserve privacy. The question is how large does Z need to be, and what distribution should it come from. To better understand this, we need a definition. Let $\Delta(D, D')$ denote the number of records that D and D' differ by, where each record corresponds to a single person.

Definition 3 (Global Sensitivity)**.** *The global sensitivity of a function f is the maximum value of the absolute difference $|f(D) - f(D')|$ over all D and D' pairs that differ in the private value of a single person. Formally:*

$$S(f) = \max_{\Delta(D,D')=1} |f(D) - f(D')|$$

The Global Sensitivity Mechanism essentially calibrates the noise added to $f(D)$ as a function of the Global Sensitivity of f. Observe that the global sensitivity is takes the maximum difference over all (D, D') pairs, and one might be tempted to replace this with the local sensitivity around the actual observed dataset D. However, the resulting local sensitivity mechanism is actually not differentially private – as the local sensitivity itself might leak some information about the dataset.

We will talk about two versions of the global sensitivity method – the Laplace Mechanism, and the Gaussian Mechanism.

The Laplace Mechanism. This mechanism adds Laplace noise that is calibrated to $S(f)/\epsilon$. Specifically, it outputs:

$$f(D) + Z, \quad Z \sim \frac{S(f)}{\epsilon}\text{Lap}(0,1)$$

where $\text{Lap}(0,1)$ is a Laplace random variable with parameters 0 and 1.

Lemma 1. *The Laplace Mechanism provides ϵ-differential privacy.*

Proof. Let $A(D)$ denote the output of the Laplace Mechanism, and let D and D' differ in the private value of a single person. For any t,

$$\begin{aligned} \frac{p(A(D)=t)}{p(A(D')=t)} &= \frac{e^{-\epsilon|t-f(D)|/S(f)}}{e^{-\epsilon|t-f(D')|/S(f)}} \\ &\leq e^{\epsilon(|t-f(D')|-|t-f(D)|)/S(f)} \\ &\leq e^{\epsilon|f(D)-f(D')|/S(f)} \leq e^{\epsilon} \end{aligned}$$

where the last step follows from the fact that $S(f) \geq |f(D) - f(D')|$. □

The Gaussian Mechanism. This mechanism adds Gaussian noise, again calibrated by a quantity proportional to $S(f)/\epsilon$. Note however that this mechanism does not lead to pure differential privacy – because the Gaussian distribution is too light-tailed. Instead we get approximate differential privacy.

In particular, the Gaussian Mechanism:

$$f(D) + Z, \quad Z \sim \mathcal{N}\left(0, \frac{2\log(1.25/\delta)S^2(f)}{\epsilon^2}\right),$$

satisfies (ϵ, δ) differential privacy. For a detailed proof, see the book by Dwork and Roth.

Example. An example of the Global Sensitivity Mechanism is in privately computing the mean of n numbers $x_1, \ldots, x_n$ that lie between 0 and 1. Since changing each number can change the sample mean by at most $1/n$, the global sensitivity is $1/n$. The Laplace Mechanism in this case will output:

$$\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n\epsilon}Z, \quad Z \sim \text{Lap}(0,1)$$

What sort of error does privacy introduce here? Roughly, the standard deviation of Z – which is the loss in accuracy due to privacy – is about $\approx \frac{1}{n\epsilon}$; this increases with lower ϵ (high privacy), and lower n (small data sizes). Observe that as we discussed earlier, this implies a privacy-accuracy-sample size tradeoff – the loss in accuracy due to privacy is lower for larger n.

Higher Dimensional Versions. So far we have been looking at the case when f is a scalar function. The global sensitivity method also extends to vector functions, but now we need to measure global sensitivity in some norm. The most commonly used version is L_2 global sensitivity, which is:

$$S_2(f) = \max_{D,D':\Delta(D,D')=1} \|f(D) - f(D')\|_2$$

The corresponding mechanisms are:

- **High Dimensional Laplace.** Output $f(D) + S_2(f)Z$, where the norm of Z is distributed as a $\Gamma(d, \epsilon)$ distribution, and the direction of Z is drawn uniformly at random.
- **High Dimensional Gaussian.** Output $f(D) + Z$, where

$$Z \sim \mathcal{N}\left(0, \frac{2\log(1.25/\delta)S^2(f)}{\epsilon^2}\right)$$

Observe that in both cases, the expected magnitude of the added noise Z is a growing function of d; it is about $\approx S_2(f) \cdot d/\epsilon$ for Laplace and about $S_2(f) \cdot \sqrt{d}/\epsilon$ for Gaussian. This is an example of what we talked about before – that the privacy-accuracy-sample-size tradeoff is worse for more complex functions.

2.4.2 The Exponential Mechanism

Another widely used differential privacy mechanism is the exponential mechanism. Given a function $f(D, x)$ where D is a sensitive dataset, and x lies in a set U, the goal of the exponential mechanism is to find an x' that approximately minimizes $f(D, x)$.

Suppose for any x, $f(\cdot, x)$ has global sensitivity C. Specifically, for all x, and all D and D' that differ by a single person,

$$|f(D, x) - f(D', x)| \leq C$$

Then, the exponential mechanism picks $x \in U$ from the following distribution:

$$p(x) \propto e^{-\epsilon f(D,x)/2C} d\mu$$

where μ is a base measure over U. We show below that this process is ϵ-differentially private.

Lemma 2. *The exponential mechanism is ϵ-differentially private.*

Proof. Let $Z(D) = \int_{x \in U} e^{-\epsilon f(D,x)/2C} d\mu(x)$. Then on input D, the exponential mechanism samples from the distribution:

$$p_D(x) = \frac{1}{Z(D)} \cdot e^{-\epsilon f(D,x)/2C} d\mu$$

If D and D' are two datasets which differ by a single person's value, then, for any x, we have:

$$\begin{aligned} \frac{p_D(x)}{p_{D'}(x)} &= \frac{Z(D')}{Z(D)} \cdot \frac{e^{-\epsilon f(D,x)/2C}}{e^{-\epsilon f(D',x)/2C}} \\ &= \frac{Z(D')}{Z(D)} \cdot e^{-\epsilon(f(D,x) - f(D',x))/2C} \leq \frac{Z(D')}{Z(D)} \cdot e^{\epsilon/2} \end{aligned}$$

where the last step follows from the fact that $|f(D,x) - f(D',x)| \leq C$. Since for each pointwise x, $e^{-\epsilon f(D,x)/2C}/e^{-\epsilon f(D',x)/2C} \leq e^{\epsilon}/2$, we have that:

$$\frac{Z(D')}{Z(D)} = \frac{\int_x e^{-\epsilon f(D',x)/2C} d\mu}{\int_x e^{-\epsilon f(D,x)/2C} d\mu} \leq e^{\epsilon/2}$$

as well. The lemma follows. □

Example. Suppose we have k classifiers $c_1, \dots, c_k$ and we are trying to evaluate which one does best on a sensitive validation dataset D. In this case, we can set $f(D, c_i)$ to be the empirical error of c_i on D. Observe that for any i, and for any two datasets D and D' which differ in a single person's private value, $|f(D, c_i) - f(D', c_i)| \leq 1/|D|$. So the exponential mechanism will sample a classifier c_i with probability:

$$p(c_i) = \frac{e^{-|D|\epsilon f(D,c_i)/2}}{\sum_{j=1}^{k} e^{-|D|\epsilon f(D,c_j)/2}}$$

Observe that again the distribution has its mode at the classifier that minimizes $f(D, c_i)$, and it gets more peaked with increasing $|D|$ (high data size) as well as high ϵ (low privacy). We thus again see this privacy-accuracy-data size tradeoff.

3 Bibliographic Notes

A formal proof that releasing too many statistics calculated based on too few people's private values can lead to a privacy breach was first shown by [DN03]; the genomic data example is due to [WLW$^+$09]. Differential privacy was introduced by [DMNS06]; the properties and the Laplace mechanism are taken from the same paper. The exponential mechanism is due to [MT07].

References

[DMNS06] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[DN03] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[MT07] Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

[WLW$^+$09] Rui Wang, Yong Fuga Li, XiaoFeng Wang, Haixu Tang, and Xiaoyong Zhou. Learning your identity and disease from research papers: information leaks in genome wide association study. In *Proceedings of the 16th ACM conference on Computer and communications security*, pages 534–544, 2009.