Statistics of Big Data, Fall 2023-24

Class notes 3

Main source for today's material: The Algorithmic Foundations of Differential Privacy by Dwork and Roth

Differential privacy definition

Assume our n sampled objects (say individuals in GWAS) belong to a finite set $\mathcal{X}$ (for GWAS $\mathcal{X}$ can be the whole population). A *sample* is now $x \in \mathbb{N}^{\mathcal{X}}$, (or more simply $x \in \{0,1\}^{\mathcal{X}}$, where $x_i = 1$ if the ith element in $\mathcal{X}$ was selected for the sample and $x_i = 0$ if not). **Remarks:**

- The typical setting is where x is a sampling indicator, so $x_i = 1$ means the sample was selected and $x_i = 0$ otherwise. However, x can have different meaning, such as a indicator of who in the population has some property, in which case $x_i = 1$ is positive and $x_i = 0$ is negative. Any setting that can be put into this indicator framework is relevant.
- Using $\mathbb{N}$ allows a situation where samples were selected more than once, for example $x_i = 3$.

The sample size is:

$$n = \|x\|_1 = \sum_{i \in \mathcal{X}} |x_i|.$$

For two samples $x, y \in \{0,1\}^{\mathcal{X}}$, denote the distance between them as the number of samples that are in one and not the other:

$$d(x,y) = \|x - y\|_1 = \sum_{i \in \mathcal{X}} |x_i - y_i|.$$

Next we define the simplex on a finite set B:

$$\Delta(B) = \left\{ p \in \mathbb{R}^{|B|} \ : \ p_i \geq 0 \ \forall i \ , \ \sum_{i=1}^{|B|} p_i = 1 \right\}.$$

Definition of a randomized algorithm: Define a *draw probability function* $M : A \to \Delta(B)$, then $\mathcal{M}$ applied to A is a random algorithm with M if:

$$\mathcal{M}(a) = b \text{ w.p. } M(a)(b), \ \forall \ a \in A, \ b \in B.$$

In words, $\mathcal{M}$ gives random output that is distributed according to M.

Now we are ready to define differential privacy:
A randomized algorithm $\mathcal{M}$ applied on $\mathbb{N}^{\mathcal{X}}$ (all possible datasets) is (ϵ, δ)-differentially private if $\forall S \subseteq Range(\mathcal{M})(= B)$, and for any $x, y \in \mathbb{N}^{\mathcal{X}}$ such that $d(x, y) \leq 1$, we have:

$$\mathbb{P}(\mathcal{M}(x) \in S) \leq e^{\epsilon} \cdot \mathbb{P}(\mathcal{M}(y) \in S) + \delta.$$

In interpreting this definition, we can see the different roles of ϵ and Δ:

- To preserve $(0, \delta)$ privacy, we can release the full information of a random portion δ of the participants, since with probability $1 - \delta$ the difference between x, y is not released. So it can lead to complete privacy violation of a small portion of the participants.
- If we preserve $(\epsilon, 0)$, it means our confidence that a specific individual is in the sample cannot change by more than $\exp(\epsilon)$ depending on the results we get reported.

Thus, it is generally considered that $\delta = 0$ called $\epsilon-$privacy is the most relevant notion, and we will not consider the case $\delta > 0$ further.

Example of ϵ-privacy preservation: Randomized response

Assume we want to ask a set of people $\mathcal{X}$ whether they do something bad (say cheat on their taxes). We instruct them to do the following:

- Flip a coin (say a fair coin, but can be a general $Ber(q)$)
- If it comes out as heads, report the true answer
- If it comes out as tails, flip another fair coin, and answer yes if it comes heads, no otherwise

Thus, 50% (or more generally, q) of the answers are true and $1 - q$ are randomly given as 50% true, and 50% false.

The statistician who analyzes the survey can easily conclude on the true percentage of cheaters via the unbiased estimate:

$$\hat{p}_{unbiased} = \frac{\hat{p} - (1 - q)/2}{q},$$

where $\hat{p}$ is the observed positive rate in the surveys.

On the other hand, this approach guarantees $\left(\log\left(\frac{1+q}{1-q}\right), 0\right)$-differential privacy. We will show it for the specific case $q = 0.5$, where $\frac{1+q}{1-q} = 3$ for simplicity of notation.

In this setting $\mathcal{X} = \{1, \ldots, n\}$, and $x \in \{0, 1\}^n$ is the identity of the true cheaters (note it is not a sampling indicator in this case). $\mathcal{M}(x)$ are the actual survey responses, and $\|x - y\| = 1$ means there is exactly one person different between x and y (cheater in one but not in the other), denote it by j and assume WLOG $x_j = 1, y_j = 0$. Since the coordinates are completely independent, and we know how the randomization works, it is easy to see that $\forall S$,:

$$\frac{\mathbb{P}\left(\mathcal{M}(x)_j = 0\right)}{\mathbb{P}\left(\mathcal{M}(y)_j = 0\right)} \leq \frac{\mathbb{P}\left(\mathcal{M}(x) \in S\right)}{\mathbb{P}\left(\mathcal{M}(y) \in S\right)} \leq \frac{\mathbb{P}\left(\mathcal{M}(x)_j = 1\right)}{\mathbb{P}\left(\mathcal{M}(y)_j = 1\right)}.$$

Given the randomization mechanism we can easily calculate:

$$\mathbb{P}\left(\mathcal{M}(x)_j = 1\right) = \frac{3}{4}\ ,\ \mathbb{P}\left(\mathcal{M}(y)_j = 1\right) = \frac{1}{4} \implies \frac{\mathbb{P}\left(\mathcal{M}(x)_j = 0\right)}{\mathbb{P}\left(\mathcal{M}(y)_j = 0\right)} = \frac{1}{3}\ ,\ \frac{\mathbb{P}\left(\mathcal{M}(x)_j = 1\right)}{\mathbb{P}\left(\mathcal{M}(y)_j = 1\right)} = 3.$$

The resulting $\log(3)-$differential privacy may not be a strong guarantee, in particular we know that a cheater is 3 times more likely to answer yes than no. What does it tell us about the probability of being a cheater given the answer is yes? Assuming the true proportion is r, we can write using Bayes rule:

$$\begin{aligned}\mathbb{P}\left(x_j = 1 | \mathcal{M}(x)_j = 1\right) &= \frac{\mathbb{P}\left(\mathcal{M}(x)_j = 1 | x_j = 1\right) \mathbb{P}\left(x_j = 1\right)}{\mathbb{P}((\mathcal{M}(x)_j = 1 | x_j = 1) \mathbb{P}\left(x_j = 1\right) + \mathbb{P}((\mathcal{M}(x)_j = 1 | x_j = 0) \mathbb{P}\left(x_j = 0\right)} = \\ &= \frac{3/4r}{3/4r + 1/4(1-r)} \leq 3r \ (\approx 3r \quad \text{if } r \text{ is small}),\end{aligned}$$

so the probability is still small, giving the person *plausible deniability.*

The Laplace Mechanism

The general idea: If I want to report some function(s) or summary(s) of the data $f(x)$, how can I "noise" it in a way that would guarantee ϵ-DP ?

Example: Counting queries. Assume we want to release K summaries on our data (in the case-control GWAS summaries example, we had $K = 10^6 \times 2$), meaning $f(x) \in \mathbb{N}^K$, where each coordinate is a count. We want to report $f(x)+r$ for some random noise $r \in \mathbb{R}^K$ that will guarantee ϵ-DP .

The Laplace mechanism is one example how to do this. Two definitions we need:

- ℓ_1 sensitivity of a function $f : \mathbb{N}^{|\mathcal{X}|} \to \mathbb{R}^K$ is:

$$\triangle f = \max_{\|x-y\|-1=1} \|f(x) - f(y)\|_1.$$

  For example, if $K = 2 \times 10^6$ and f are counts, it is easy to see that $\triangle f = K = 2 \times 10^6$, for example if the observation that is in x and not in y has all the counted properties turned on.

- The Laplace distribution (AKA double exponential distribution) is a continuous distribution denoted $X \sim Lap(b)$, with density:

$$p(x) = \frac{1}{2b} \exp\left(-\frac{|x|}{b}\right).$$

  It is symmetric distribution around zero with $\mathbb{E}(X) = 0,\ Var(X) = \mathbb{E}(X^2) = 2 \cdot b^2$.

Laplace Mechanism definition: Given a K dimensional release problem as above, draw random variables

$$Y_1, \dots Y_K \sim Lap\left(\frac{\triangle f}{\epsilon}\right) \quad \text{i.i.d},$$

and report the randomized summaries:

$$\mathcal{M}_{Lap,f,\epsilon}(x)_j = f(x)_j + Y_j \ \ , \ j = 1, \dots, K.$$

Theorem: The Laplace mechanism $\mathcal{M}_{Lap,f,\epsilon}$ guarantees ϵ-DP .

Proof: Take $x, y \in \mathbb{N}^{|\mathcal{X}|}$ two datasets with $\|x - y\|_1 \leq 1$. We denote $\mathcal{M}_{Lap,f,\epsilon}(x) \sim p_x$ the density function when the truth is x and similarly p_y. Take a point $z \in Range(\mathcal{M}) \subseteq \mathbb{R}^K$, and check the ratio of densities under the two distributions:

$$\begin{aligned} \tfrac{p_x(z)}{p_y(z)} &= \prod_{k=1}^{K} \frac{\exp\left(-\frac{\epsilon}{\triangle f}|f(x)_k - z_k|\right)}{\exp\left(-\frac{\epsilon}{\triangle f}|f(y)_k - z_k|\right)} = \prod_{k=1}^{K} \exp\left(\frac{\epsilon}{\triangle f}|f(y)_k - z_k| - |f(x)_k - z_k|\right) \\ &\overset{(*)}{\leq} \prod_{k=1}^{K} \exp\left(\frac{\epsilon}{\triangle f}|f(y)_k - f(x)_k|\right) = \exp\left(\frac{\epsilon}{\triangle f}\sum_k |f(y)_k - f(x)_k|\right) = \exp\left(\frac{\epsilon}{\triangle f}\|f(y) - f(x)\|_1\right) \overset{(**)}{\leq} e^{\epsilon}, \end{aligned}$$

where $(*)$ is due to properties of absolute value (same if the differences have $+, -$ signs, bigger in all other cases), and $(**)$ is due to the definition of $\triangle f$.
This bound on the ratio holds for all z and therefore also for any group of such z values $S \subseteq Range(\mathcal{M})$.

Conclusion: If we want K counts from GWAS we need to noise each one with noise $Lap\left(\frac{K}{\epsilon}\right)$, whose variance is $2K^2/\epsilon^2$. For $K = 2 \cdot 10^6, \epsilon = \log 2$, the standard deviation of the noise is therefore:

$$sd = \frac{2\sqrt{2} \cdot 10^6}{\log 2} \approx 4 \cdot 10^6.$$

This means we can release the number of carriers of each property to within about $\pm 4 \cdot 10^6$ (note this is independent of the sample size n).
$\Rightarrow$ unless the sample size of individuals is in the many millions, this is not a useful way to release that many summaries.

How can we make it useful? If we wanted to release a much smaller number of summaries then that would obviously change: for $K = 10$ we easily see that the Laplace noise standard deviation will be only 20, meaning if $f(x)$ is in the thousands, the information will be useful.

It is also important to note that the Laplace mechanism (like other mechanisms we will discuss) is *sufficient* for ϵ-DP but not necessary, so it may be very suboptimal.

An interesting example: Report noisy max. Assume that we only want to know which of the K counts or summaries is the biggest (e.g. most common disease in a health dataset). Naive release with Laplace mechanism requires $Lap(K/\epsilon)$ noise, which is deadly for large K as we showed. The book shows that in this case it is enough to do the following:

- Add noise of $Lap(1/\epsilon)$ to each of the K counts, regardless of K — this has standard deviation of only $\sqrt{2}/\epsilon$.
- Report the index $k \in \{1, \ldots K\}$ of the biggest noisy count (not the count itself, or any of the counts).

The information we gain is limited (which count is largest) but useful and likely correct.

The exponential mechanism

Assume now we also have a *utility function:*

$$u : \mathbb{N}^{\mathcal{X}} \times B \longrightarrow \mathbb{R},$$

where $u(x,l)$ is a measure of how much utility we get out of reporting l when the true data is x. For example, if we want to report the average of the data $\bar{x}$, the utility might be:

$$u(x,l) = -\|\bar{x} - l\|_q^q,$$

where $q=1$ gives absolute error and $q=2$ squared error. This mechanism allows us to combine adding noise to preserve privacy with not hurting the utility too much and preserving the "relevant" information.

As before, define the sensitivity:

$$\triangle_u = \max_{l \in B} \max_{\|x-y\|_1 \le 1} |u(x,l) - u(y,l)|,$$

the maximal possible difference in utility of the same reported result between neighbors.

Now given x we want our randomized algorithm to prefer l's for which the utility $l(x,u)$ is high, unlike in Laplace where we added completely random noise. Therefore we will give higher probability to high utility outcomes, specifically the exponential mechanism with utility u and privacy parameter ϵ, denoted $\mathcal{M}_{E,u,\epsilon}$ uses the following distribution:

$$\mathbb{P}(\mathcal{M}_{E,u,\epsilon}(x) = l) \propto \exp\left\{\frac{\epsilon u(x,l)}{2\triangle_u}\right\}.$$

(Note it is proportional and not equal since the quantity on the right is generally not a distribution.
Theorem: The exponential mechanism $M_{E,u,\epsilon}$ preserves ϵ-DP for any mechanism u.
Intuition of proof: If the quantity on the right was indeed a distribution, i.e.:

$$\mathbb{P}(\mathcal{M}_{E,u,\epsilon}(x) = l) = \exp\left\{\frac{\epsilon u(x,l)}{2\triangle_u}\right\},$$

then we would have:

$$\log\left(\frac{\mathbb{P}(\mathcal{M}(x) = l)}{\mathbb{P}(\mathcal{M}(y) = l)}\right) = \epsilon\frac{u(x,l) - u(y,l)}{2\triangle} \le \frac{\epsilon}{2},$$

and we would have $\epsilon/2$-DP. Since it is not equal but proportional, both sides have to be divided by the sums over l, and using the definition of $\triangle_u$ again gives the other $\epsilon/2$.
Example: reporting the mean. Assume our data $x = (X_1, \dots X_n)$ is an iid sample of size n from some distribution F and we want to report the mean $f(x) = \bar{X}$ as an estimate of $\mu = \mathbb{E}F$ in a private manner. Assume also the support of F is finite, say $X_i \in [0,1]$. We could use the Laplace mechanism, it is easy to see:

$$\triangle f = \frac{1}{n} \;\Rightarrow\; \mathcal{M}_{L,\epsilon}(x) = \bar{X} + Lap(\frac{1}{n\epsilon}) \;\Rightarrow\; \mathbb{P}(\mathcal{M}_{L,\epsilon}(x) = l) \propto \exp\left\{-n \cdot \epsilon \cdot |\bar{X} - l|\right\}.$$

On the other hand, we could apply the exponential mechanism with $u(x,l) = -|\bar{X} - l|$, then we get:

$$\triangle_u = \max_{l} \max_{\|x-y\|_1 = 1} \big|\, |\bar{X} - l| - |\bar{Y} - l| \,\big| \le \max_{\|x-y\|_1 = 1} |\bar{X} - \bar{Y}| = \frac{1}{n},$$

and we get a similar but slightly worse result that:

$$\mathbb{P}(\mathcal{M}_{L,\epsilon}(x) = l) \propto \exp\left\{-\frac{n \cdot \epsilon}{2} \cdot |\bar{X} - l|\right\},$$

equivalent to adding $Lap(2/(n\epsilon))$ with bigger variance.

A more interesting application of the exponential mechanism would use $u(x,l) = -(\bar{X} - l)^2$ the Euclidean distance. In this case we can similarly show that $\delta_u \leq 1/n$ and therefore the exponential mechanism would give:

$$\mathbb{P}(\mathcal{M}_{L,\epsilon}(x) = l) \propto \exp\left\{-\frac{n \cdot \epsilon}{2} \cdot (\bar{X} - l)^2\right\},$$

meaning we know that it has a normal distribution:

$$l|x \sim N(\bar{X}, \frac{1}{n\epsilon}) \quad \Rightarrow \quad l \sim N\left(\mu, \frac{1}{n}\left(\frac{1}{\epsilon} + \sigma^2\right)\right),$$

where the last step shows the unconditional distribution of l as an estimate of μ.
We therefore conclude that $l = \mu + O_p(1/\sqrt(n))$, so the convergence rate of l to μ is the same as that of the average $\bar{X}$, even if its variance is bigger.