Statistics of Big Data, Fall 2021-22

Class notes 4

Sources for today's material:
The Algorithmic Foundations of Differential Privacy by Dwork and Roth
A Statistical Framework for Differential Privacy by Wasserman and Zhou

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:

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

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 and 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:
$$\triangle_u=\max_{l}\max_{\|x-y\|_1=1}|\;|\bar{X}-l|-|\bar{Y}-l|\;|\le\max_{\|x-y\|_1=1}|\bar{X}-\bar{Y}|=\frac{1}{n},$$
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 $\delta_u\le 1/n$ 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.

Example from Wasserman and Zhu: Estimating the full density while preserving DP

Assume as before our data is $x=(X_1,...X_n)$ with $X_i\sim F$ iid, now a multivariate distribution with $supp(F)\subseteq[0,1]^r$. We want to estimate the entire density of F.

The obvious non-private estimate is a *histogram:* Divide $[0,1]^r$ into $m=V^r$ regions (each dimension divided into V bins, say equal sized), and report a vector of length m of counts, denoted $\hat{F}_x$. However we want to report F_x while maintaining ϵ-DP . The Laplace mechanism would require adding $Lap(m\cdot\triangle f/\epsilon)$ noise, and since m is exponential this can be huge and destroy the usefulness.

Instead, Wasserman and Zhou suggest using the exponential mechanism for releasing a "sample" of length k: $z=(Z_1,\dots,Z_k)$ that is "similar" to $\hat{F}_x$ but private. Denote the empirical distribution of Z by $\hat{F}_Z$, they propose to use distance between the distributions in the mechanism, for example the Kolmogorov-Smirnov (KS) distance:

$$u(x,z) = KS(\hat{F}_x, \hat{F}_z) = \max_{y\in[0,1]^r} |\hat{F}_x - \hat{F}_z|.$$

(Reminder: the KS distance of two one-dimensional cumulative distribution functions F, G is $\sup_{x\in\mathbb{R}} |F(x) - G(x)|$, with a natural extension for multivariate cumulative distribution functions). For this KS-based utility it is easy to verify that $\triangle_u = 1/n$, and therefore every set z should be drawn with probability distribution:

$$h(z|x) \propto \exp\left\{-\frac{\epsilon \cdot n \cdot KS(\hat{F}_x, \hat{F}_z)}{2}\right\},$$

and would preserve ϵ-DP .

Wasserman and Zhou compare this mechanism to adding noise to the histogram (as in the Laplace proposal above) or pertrubing the histogram in other ways that preserve ϵ-DP , and conclude that this exponential mechanism is the best in the sense of convergence of the reported histogram $\hat{F}_z$ to the true distribution F :

Theorem (W&Z): If we choose $k(n)$ to be big enough and use the KS exponential mechanism above, then:

$$KS(F, \hat{F}_z) = O_p\left(n^{-\frac{1}{3}}\right).$$

This rate is faster than the other ϵ-DP methods above, although in this case it is slower than reporting the noiseless histogram:

$$KS(F, \hat{F}_x) = O_p\left(n^{-\frac{1}{2}}\right).$$

Practical problem: Recall that this u function depends on the entire vector z of k observations, it is not clear from the paper how such vectors can be practically drawn, whereas the less efficient noising or pertrubation methods can be practically implemented ⇒ requires further research.

W&Z have plenty of other interesting results and analyses about the use of ϵ-DP in statistical methodology, here is a very simple one which demonstrates the obvious fact that ϵ-DP destroys information on each specific subject in our data, and therefore precludes making confident conclusions about them (good for privacy, bad for statistical inference!): **Theorem (W&Z):** Assume again our data x is an iid sample of size n: $X_i \sim F$ and assume *we know* F. We have an ϵ-DP mechanism $\mathcal{M}$ (and obviously we also know the distributions m underlying it). Now we observe released data $\mathcal{M}(x)$, then any test of $H_0: \; X_i = U$ vs $H_A: \; X_i = V$ at level α has most power $\alpha \exp(\epsilon)$. Hence, not surprisingly it is impossible to obtain substantial power for testing whether a specific observation is in the dataset.

Summary

Both our sources, especially the mini-book by Dwork and Roth (D& R) have plenty more results, for example D& R describe *composition rules* which demonstrate that if we have a sequence of noising mechanisms:

$$x \to \mathcal{M}_1(x) \to \mathcal{M}_2(\mathcal{M}_1(x)),$$

and the first mechanism is ϵ_1-DP, while the second is ϵ_2-DP, then the combined mechanism which releases only $\mathcal{M}_2(\mathcal{M}_1(x))$ is $\epsilon_1 \cdot \epsilon_2$-DP, which can be very useful in practice.

We have only tasted some of the definitions and methods involved this area, and in reading Z&W we also tried to get a better understanding of the meaning of these results in the context of statistical problems like estimation and hypothesis testing.

Summarizing our discussion of differential privacy:

1. Privacy is a real and important problem (see our discussion of GWAS and genetic information release)

2. Differential privacy is theoretically elegant but often very conservative. One reason is the requirement that the guarantees hold for every possible $S \subseteq B$.

3. The known ϵ-DP solutions lead to methods that are not practical for important problems like GWAS. Note however that all we have are *sufficient* solutions which maintain ϵ-DP and not *optimal* methods in provable senses. Also for problems of low dimension like releasing a single average or noisy-max, the ϵ-DP solutions can be acceptable.

4. The basic approach is very "computer-science" oriented, for example it does not assume any distribution over the population or for the sampling from $\mathbb{N}^{\mathcal{X}}$, rather it requires that the results hold for all possible values, and the only randomness comes from the reporting mechanism $\mathcal{M}$. Results like those of W&Z add a level by assuming iid sampling from a distribution and asking how good the results are as statistical estimates of the population quantities.

5. The mathematical thinking behind ϵ-DP is useful for other statistical applications like measuring and avoiding overfitting in repeatedly using the same data for estimation. We will discuss this later in the course in the context of dealing with adaptive data analysis.