

## 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**

### Summary

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

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

and the first mechanism is  $\epsilon_1$ -DP, while the second is  $\epsilon_2$ -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.

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  $\epsilon$ -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  $\epsilon$ -DP and not *optimal* methods in provable senses. Also for problems of low dimension like releasing a single average or noisy-max, the  $\epsilon$ -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 that we discuss later today 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  $\epsilon$ -DP is useful for other statistical applications like measuring and avoiding overfitting in repeatedly using the same data for estimation. We may discuss this later in the course in the context of dealing with adaptive data analysis, as time permits.

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

Assume as before our data is  $x = (X_1, \dots, X_n)$  with  $X_i \sim F$  iid, now a multivariate distribution with  $\text{supp}(F) \subseteq [0, 1]^r$ . 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  $\epsilon$ -DP. The Laplace mechanism would require adding  $\text{Lap}(m \cdot \Delta f / \epsilon)$  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  $\Delta_u = 1/n$ , 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  $\epsilon$ -DP.

Wasserman and Zhou compare this mechanism to adding noise to the histogram (as in the Laplace proposal above) or perturbing the histogram in other ways that preserve  $\epsilon$ -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  $\epsilon$ -DP methods above when the dimension  $r \geq 5$ , 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 perturbation methods can be practically implemented  $\Rightarrow$  requires further research.

W&Z have plenty of other interesting results and analyses about the use of  $\epsilon$ -DP in statistical methodology, here is a very simple one which demonstrates the obvious fact that  $\epsilon$ -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  $\epsilon$ -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  $\alpha$  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.