CSC2419:Algorithms and Complexity in Private Data Analysis Winter 2018

Lecture 11: Adaptive Data Analysis via DP

Aleksandar Nikolov *Scribe: James Lucas*

1 Set up

Assume we are given some unknown distribution $\mathcal{D}$ on $\mathcal{X}$. We want to know what fraction of the population defined by $\mathcal{D}$ has some property. Formally, we define,

$$q_1, \dots, q_m : \mathcal{X} \to \{0, 1\}$$

We want to estimate,

$$q_i(\mathcal{D}) := \mathbb{E}_{x \sim \mathcal{D}}[q_i(x)], \quad \forall i$$

The classical solution to this problem is as follows. Let $X = (x_1, \dots, x_n)$, where every x_i is drawn independently from $\mathcal{D}$. We estimate the true mean $q_i(\mathcal{D})$ by the empirical mean,

$$q_i(X) = \frac{1}{n}\sum_{j=1}^{n} q_i(x_j)$$

.

To make sure that these are good estimates, we need to know how close the empirical means are to the true means, $q_i(D)$. Note that $\mathbb{E}[q_i(X)] = q_i(\mathcal{D})$. We can compute a bound using Hoeffding's inequality:

$$\begin{aligned} \forall i: \quad & \mathbb{P}(|q_i(X) - q_i(D)| > \alpha) \le 2\exp\left(-\frac{2\alpha^2}{n}\right) \\ \Rightarrow & \mathbb{P}(\exists i : |q_i(X) - q_i(D)| > \alpha) \le 2m\exp\left(-\frac{2\alpha^2}{n}\right) \le \beta \end{aligned}$$

To satisfy this, we need

$$n \ge \frac{\log(2m/\beta)}{2\alpha^2}$$

This reasoning is, however, no longer valid if q_i is based on $q_1(X), \dots, q_{i-1}(X)$? (We refer to these as *adaptive* queries). In Hoeffding's inequality, we assume that each element of the sum (in the empirical mean) is independent and lies in the interval [0,1]. Moreover, if q_i depends on the data, we may no longer have an unbiased estimator for the mean. In short: things can break!

2 Artificial example

When $m \gg n$, then, under some assumptions on the q_i, we can recover X from $q_1(X), \ldots, q_{m-1}(X)$. Then we can define our next query in the following way,

$$q_m(x') = \begin{cases} 1, & \text{if } x_i = x' \text{ for some } i, \\ 0, & \text{otherwise.} \end{cases}$$

Note that $q_m(X) = 1$. If $\mathcal{D}$ is uniform on $\mathcal{X}$, then, for $|\mathcal{X}| = n^{100}$, we have $q_m(\mathcal{D}) = \frac{1}{n^{99}} \approx 0$. This can be thought of as an example of catastrophic overfitting to the data!

3 Naive solution and Improvement

A simple way to handle adaptive queries is to partition X into $X^1, \ldots, X^m$ and use X^i to compute $q_i(X) \approx q_i(\mathcal{D})$. For this naive solution, we would need,

$$n \geq \frac{m \log(1/\beta)}{\alpha^2}.$$

This is a lot more data then before. Can we do better? We will show that using differential privacy we can! In particular, we will show that we can use,

$$n \gtrsim \frac{\sqrt{m \log(2m/\beta)}}{\alpha^2}.$$

Roughly speaking, this is the best possible bound. It is achieved via the following Transfer Theorem.

Theorem 1 *(**Transfer Theorem**) Suppose that the mechanism $\mathcal{M}$ takes n iid samples X from $\mathcal{D}$, and answers m (adaptive) queries $q_1, \ldots, q_m$ on X such that,*

i) $\forall X \in \mathcal{X}^n, \mathbb{P}_{\mathcal{M}}(\exists i : |q_i(X) - q_i(D)| > \alpha) \leq \beta$

ii) $\mathcal{M}$ *is* $(\alpha, \alpha\beta)$*-DP.*

Then,

$$\mathbb{P}_{\mathcal{D},\mathcal{M}}(\exists i : |q_i(X) - q_i(D)| > C\alpha) \leq C\beta, \tag{1}$$

for some constant C.

Proof: See, e.g. Bassily et al. [2016]. We prove a slightly weaker version below. ■

By adaptive, we mean that q_i can depend on $\mathcal{M}$'s answers to $q_1, \ldots, q_{i-1}$. Intuitively, adaptive data queries may be able to overfit to the data. Differential privacy corrects this as queries are only allowed to receive aggregate information on the data - making overfitting more difficult. By instantiating the Transfer Theorem with mechanisms we have seen in this class, we can get different sample bounds that improve on the naive soluion.

Gaussian noise mechanism: We can directly plug this in to get the bound from Equation 1.

Online PMW: For this mechanism, we need:

$$n \gtrsim \frac{\sqrt{\log |\mathcal{X}| \log(m/\beta) \log(1/\alpha\beta)}}{\alpha^3}.$$

4 Main Lemma and Generalization in Machine Learning

The proof of the Transfer Theorem relies on the following lemma.

Lemma 2 *Suppose $\mathcal{W}$ is (ϵ, δ)-DP and on input $X \in \mathcal{X}^n$, it outputs a counting query q. Let $X \sim \mathcal{D}^n$ (independent rows). Then,*

$$|\mathbb{E}_{X,\mathcal{W}}[q(D)|q = \mathcal{W}(X)] - \mathbb{E}_{X,\mathcal{W}}[q(X)|q = \mathcal{W}(X)]| \leq e^\epsilon - 1 + \delta.$$

Note that the conditioning above simply defines the random query q. We use it only for clarity. In words, the lemma states that a differentially private algorithm cannot do what we saw in the artificial example before, i.e. cannot find a query which distinguishes the data from the underlying distribution.

Before we prove the lemma, let us see how it immediately implies that differentially private empirical risk minimization generalizes to unseen samples. Let us recall the basic set up for supervised learning. We define the 0-1 loss as,

$$l(y', y) = \begin{cases} 1, & y' \neq y, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

We also define the loss on $\mathcal{D}$ as,

$$L(\theta, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[l(f_\theta(x_i), y_i)],$$

and the empirical loss as

$$L(\theta, X) = \frac{1}{n}\sum_{i=1}^{n} l(f_\theta(x_i), y_i).$$

Here $\{f_theta : \theta \in \Theta\}$ is the hypothesis class, defined as functions parametrized by some parameter vector θ.

Notice that the empirical loss is a counting query for every θ. Then Lemma 2 implies that if θ is computed on X by an (ϵ, δ)-DP algorithm $\mathcal{M}$, then,

$$\mathbb{E}_{X,\mathcal{M}}[L(\theta, \mathcal{D})] \leq \mathbb{E}_{X,\mathcal{M}}[L(\theta, X)] + e^\epsilon - 1 + \delta.$$

In fact, you can check that our proof of the lemma actually shows that for any non-negative loss l, with values in $[0, L]$,

$$\mathbb{E}_{X,\mathcal{M}}[L(\theta, \mathcal{D})] \leq e^{\epsilon}\mathbb{E}_{X,\mathcal{M}}[L(\theta, X)] + \delta L.$$

Thus the loss achieved by differentially private empirical risk minimization on the data is never much more than the loss achieved on the true distribution.

Next we prove the lemma.

Proof: For brevity we will not include the $X, \mathcal{M}$ subscript from the expectations, with the understanding that all expectations and probabilities are with respect to the randomness of both X and $\mathcal{M}$. By linearity of expectation, and because $q(x_i) \in \{0, 1\}$

$$\begin{aligned}\mathbb{E}[q(X)|q = \mathcal{W}(X)] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[q(x_i)|q = \mathcal{W}(X)] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(q(x_i) = 1|q = \mathcal{W}(X)).\end{aligned}$$

We now apply differential privacy. Take $x_i' \sim \mathcal{D}$ independent of all else. Then define $X' = (x_1, \ldots, x_i', \ldots, x_n)$ such that X and X' are neighboring. Then, by the definitions of differential privacy,

$$\mathbb{P}(q(x_i) = 1|q = \mathcal{W}(X)) \leq e^{\epsilon}\mathbb{P}(q(x_i) = 1|q = \mathcal{W}(X')). \tag{3}$$

Notice that the joint distribution of (x_i, X') is identical to the joint distribution of (x_i', X): in either case we have an item sampled from $\mathcal{D}$ and n other items, independently sampled from $\mathcal{D}$. Then we must have,

$$\begin{aligned}\mathbb{P}(q(x_i) = 1|q = \mathcal{W}(X')) &= \mathbb{P}(q(x_i') = 1|q = \mathcal{W}(X)) \\ &= \mathbb{E}[q(D)|q = \mathcal{W}(X)]\end{aligned}$$

Substituting this back in, we get,

$$\mathbb{E}[q(x)|q = \mathcal{W}(X)] \leq e^{\epsilon}\mathbb{E}[q(D)|q = \mathcal{W}(X)] + \delta$$

This is not quite what we wanted to prove. Notice that $A - B \leq (e^{\epsilon} - 1)B + \delta \leq e^{\epsilon} - 1 + \delta$. We can repeat the previous argument from Equation 3 in reverse and combine the bounds in this way to recover the original statement (left as an easy exercise). ■

5 A Transfer Theorem

We show the following, somewhat easier, transfer theorem.

Theorem 3 (Easier Transfer Theorem) *Let $X \sim \mathcal{D}^n$ and let $\mathcal{M}$ be (ϵ, δ)-DP such that for every adaptive $q_1, \ldots, q_m$ and for all $X \in \mathcal{X}^n$,*

$$\mathbb{E}_{\mathcal{M},X}[\max_i |q_i(X) - \mathcal{M}(X)_i|] \leq \alpha.$$

Then,

$$\mathbb{E}_{\mathcal{M},X}[\max_i |q_i(\mathcal{D}) - \mathcal{M}(X)_i|] \leq \alpha + e^\epsilon - 1 + \delta.$$

Proof: Take $\mathcal{W}$ which simulates $\mathcal{M}$ on $q_1, \ldots, q_m$ and outputs q_i that maximizes $|q_i(\mathcal{D}) - \mathcal{M}(X)_i|$. Then $\mathcal{W}$ is (ϵ, δ)-DP, as we can think of it as a post-processing of $\mathcal{M}$ that does not use the dataset X.

We use the following trick: let us change the queries and the mechanism so that, whenever q_i is asked, then $1 - q_i$ is also asked, and is answered by $1 - \mathcal{M}(X)_i$. Then,

$$\max_i |q_i(D) - \mathcal{M}(X)_i| = \max_i q_i(D) - \mathcal{M}(X)_i,$$

which leads to,

$$\begin{aligned}\mathbb{E}_{X,\mathcal{M}}[\max_j q_j(D) - \mathcal{M}(X)_j] &= \mathbb{E}[q_i(D) - \mathcal{M}(X)_i | q_i = \mathcal{W}(X)] \\ &= \mathbb{E}[q_i(D) - q_i(X) | q_i = \mathcal{W}(X)] + \mathbb{E}[q_i(X) - \mathcal{M}(X)_i | q_i = \mathcal{W}(X)]\end{aligned}$$

The first expectation is at most $e^\epsilon - 1 + \delta$ and the second is at most α (by assumption). ■

The only difference with the first transfer theorem is that this one does not give high-probability bounds on the errors, but rather gives a bound on the expected worst-case error.

References

Raef Bassily, Kobbi Nissim, Adam D. Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *STOC*, pages 1046–1059. ACM, 2016.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.