CSC2412: Adaptive Data Analysis via Differential Privacy

Sasho Nikolov

The adaptive data analysis problem

Estimating population counts

universe of possible data points

- Unknown distribution D on $\mathcal{X}$

↪ models the population

- Predicates $q_1, \ldots, q_k : \mathcal{X} \to \{0, 1\}$

E.g. q_1 = ? smoker
q_2 = ? smoker and male
q_3 = ? smoker and PhD
...

Want to estimate, for all $i = 1 \mathinner{..} k$:

$$q_i(D) = \mathbb{E}_{x \sim D}[q_i(x)].$$

↪ fraction of the population satisfying q_i

The classical solution

Draw a sample $X = \{x_1, \ldots, x_n\}$ iid from D.

Hope that $\forall i : q_i(X) \approx q_i(D)$

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

↳ independent, in $\{q_i\}$

$$\mathbb{E}[q_i(X)] = q_i(D)$$

Hoeffding: $\forall i \quad \mathbb{P}(|q_i(X) - q_i(D)| > \alpha) = \mathbb{P}(|q_i(X) - \mathbb{E}q_i(X)| > \alpha)$

$$\leq 2e^{-2n\alpha^2}$$

$$\mathbb{P}(\exists i : |q_i(X) - q_i(D)| > \alpha) \leq 2k \cdot e^{-2n\alpha^2} \leq \beta \quad \text{if} \quad \boxed{n \geq \frac{\ln(2k/\beta)}{2\alpha^2}}$$

Adaptive queries?

What if q_i depends on ~~$q_1, \ldots, q_{i-1}$~~? the estimates for $q_1(D), \ldots, q_{i-1}(D)$

E.g. q_i is chosen based on $q_1(X), \ldots, q_{i-1}(X)$

E.g. $q_1 =$? smokers and male
$q_2 =$? smokers and female $\Big\} \rightarrow$ if even split ask $q_3 =$? smokers and ≥ 35 yrs

else stop

Suppose we ask $q_1(X), q_2(X), \ldots, q_k(X)$ for $k \gg n$, q_i random predicates and we "invert" to learn X

$$q_{k+1}(x) = \begin{cases} 1 & x \in X \\ 0 & \text{o/w} \end{cases} \Rightarrow q_{k+1}(X) = 1. \text{ But if } D \text{ is uniform on } \mathcal{X} \text{ then } q_{k+1}(D) \approx 0$$

A simple solution

Break $X = \{x_1, \ldots, x_n\}$ into $X^1 = \{x_1, \ldots, x_{n/k}\}$

$X^2 = \{x_{n/k+1}, \ldots, x_{2n/k}\}$

$\vdots$

$X^k = \{x_{\frac{(k-1)n}{k}+1}, \ldots, x_n\}$

Answer $q_1(D)$ by $q_1(X^1)$

q_2 by $q_2(X^2)$

$\vdots$

q_k by $q_k(X^k)$

to get error α w/ prob $1-\beta$

I need $\frac{n}{k} \geq \frac{\ln(2k/\beta)}{2\alpha^2} \iff \boxed{n \geq \frac{k \ln(2k/\beta)}{2\alpha^2}}$

Can we do better?

Transfer theorem

$\mathcal{M}$ answers q_1 w/ $\mathcal{M}(X)_1$

q_2 determined from $\mathcal{M}(X)_1$ $\rightarrow$ $\mathcal{M}$ answers w/ $\mathcal{M}(X)_2$
by analyst

....

Theorem
Suppose $\mathcal{M}$ takes a dataset X and answers k adaptive queries $q_1, \dots, q_k$. If

1. $\forall X \in \mathcal{X}^n, \mathbb{P}(\exists i : |q_i(X) - \mathcal{M}(X)_i| > \alpha) < \alpha\beta,$ $\rightarrow$ $\mathcal{M}$ accurate on the dataset
2. $\mathcal{M}$ *is* $(\alpha, \alpha\beta)$*-DP,*

then for a constant C

$$\mathbb{P}_{\substack{\mathcal{M} \\ X \sim D^n}}(\exists i : |\mathcal{M}(X)_i - q_i(D)| > C\alpha) < C\beta.$$

$X \sim D^n \iff X = \{x_1, \dots, x_n\}$ $x_i \sim D$ independently

Improving on the simple solution

Simple solution: error α with $\approx \frac{k \log(k/\beta)}{\alpha^2}$

Can get error α with $\approx \frac{\sqrt{k \log k}}{\alpha^2}$ samples.

Gaussian noise + advanced composition

answer q_i w/ $q_i(X) + Z_i$ $\quad Z_i \sim N\left(0, \boxed{\frac{1}{n^2 \cdot \rho}}\right)$ $\approx \frac{k \log 1/\delta}{n^2 \alpha^2}$

and we get (ε, δ)-DP

for any δ and $\varepsilon \approx \sqrt{k \rho \ln(1/\delta)}$

Transfer thm: we need $(\alpha, \alpha\beta)$-DP $\quad \rho \approx \frac{\alpha^2}{k \log(1/\delta)}$

std dev per q_i is $\approx \frac{\sqrt{k \ln 1/\delta}}{n \alpha} = \frac{\sqrt{k \ln(1/\alpha\beta)}}{n \alpha} \ll \alpha$ if $n \gg \frac{\sqrt{k \ln(1/\alpha\beta)}}{\alpha^2}$

Key Lemma

$X \in \mathcal{X}^n$

$q : \mathcal{X} \to \{0,1\}$

D is a distr. on $\mathcal{X}$

Lemma

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

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

over random choice of $X \sim D^n$
and randomness of $\mathcal{W}$

A DP algorithm cannot find a query that distinguishes X from D.

Proof of Key Lemma

$q(X) = \frac{1}{n}\sum_{i=1}^{n} q(x_i)$ $\qquad q : \mathcal{X} \to \{0,1\}$

$$\mathbb{E}[q(X) \mid q = \mathcal{W}(X)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[q(x_i) \mid q = \mathcal{W}(X)] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}(q(x_i) = 1 \mid q = \mathcal{W}(X))$$

Take $x_i' \sim D$ independently from everything else.

$X' = \{x_1, \dots, x_{i-1}, x_i', x_{i+1}, \dots, x_n\}$ $\qquad$ X, X' neighbouring

$$\mathbb{P}(q(x_i) = 1 \mid q = \mathcal{W}(X)) \boxed{\le} e^{\varepsilon}\, \mathbb{P}(q(x_i) = 1 \mid q = \mathcal{W}(X')) + \delta$$

(ε, δ)-DP of $\mathcal{W}$

Proof part 2

$X = \{x_1, \dots, x_n\}$

$X' = \{x_1 \dots, x_i', x_{i+1}, \dots, x_n\}$

Observation: (x_i, X') has the same distribution as (x_i', X)

$$\mathbb{P}(q(x_i)=1 \mid q = \mathcal{W}(X)) \le e^{\varepsilon}\, \mathbb{P}(q(x_i)=1 \mid q = \mathcal{W}(X')) + \delta$$

$$= e^{\varepsilon}\, \mathbb{P}(q(x_i')=1 \mid q = \mathcal{W}(X)) + \delta$$

$$= e^{\varepsilon}\, \mathbb{E}\{q(D) \mid q = \mathcal{W}(X)\} + \delta$$

$$q(D) = \mathbb{E}_{x\sim D}\, q(x) = \mathbb{P}_{x\sim D}(q(x)=1)$$

$$\mathbb{E}[q(X) \mid q = \mathcal{W}(X)] \le e^{\varepsilon} \underbrace{\mathbb{E}[q(D) \mid q = \mathcal{W}(X)]}_{\le 1} + \delta$$

$$\mathbb{E}\{q(X) \mid q = \mathcal{W}(X)\} - \mathbb{E}\{q(D) \mid q = \mathcal{W}(X)\} \le e^{\varepsilon} - 1 + \delta$$

$$\ge -(e^{\varepsilon} - 1 + \delta) \quad \text{analogous}$$

Aside: Generalization from DP

Almost the same proof as the lemma (exercise)

Theorem

For any non-negative loss $\ell(\theta,(x,y))$, $X = \{(x_1,y_1),\ldots,(x_n,y_n)\} \sim D^n$, and

$$L_X(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(\theta,(x_i,y_i)) \qquad L_D(\theta) = \mathbb{E}_{(x,y)\sim D}[\ell(\theta,(x,y))],$$

if θ is computed by an (ε,δ)-DP algorithm, then

$$\mathbb{E}[L_D(\theta)] \le e^{\varepsilon}\mathbb{E}[L_X(\theta)] + \delta \max_{\theta,x,y} \ell(\theta,(x,y)).$$

DP implies generalization

Population loss is not much more than empirical loss for DP algo.

A simpler transference theorem

Theorem
If the mechanism $\mathcal{M}$ satisfies that

1. $\forall X \in \mathcal{X}^n$, *and all sequence of adaptive queries* $q_1, \ldots, q_k$,
   $\mathbb{E}[\max_i |q_i(X) - \mathcal{M}(X)_i|] \leq \alpha$
   $1-q_1 \qquad 1-q_k$
2. $\mathcal{M}$ *is* (ε, δ)*-DP,*

then

$$\mathbb{E}[\max_i |q_i(D) - \mathcal{M}(X)_i|] \leq \underline{\alpha + e^{\varepsilon} - 1 + \delta} \approx \alpha + \varepsilon + \delta$$

$q_1, \ldots, q_k$ are adaptively chosen based on $\mathcal{M}(X)_1, \ldots, \mathcal{M}(X)_k$

$X \sim D^n$

Proof

$q_i(x) = 1 \iff 1 - q_i(x) = 0$
$q_j(x) = 0 \iff 1 - q_i(x) = 1$

Trick: Suppose that if q_i is asked, so is $1 - q_i$, and is answered by $1 - \mathcal{M}(X)_i$.

Then $\max_{i=1}^{k} |q_i(D) - \mathcal{M}(X)_i| = \max_{i=1}^{k} q_i(D) - \mathcal{M}(X)_i$.

$$|q_i(D) - \mathcal{M}(X)_i| = \max\{q_i(D) - \mathcal{M}(X)_i, \underbrace{\mathcal{M}(X)_i - q_i(D)}_{= 1 - q_i(D) - (1 - \mathcal{M}(X)_i)}\}$$

Define $\mathcal{W}$ s.t. it
*) Simulates $\mathcal{M}$ on the adaptive queries $q_1, \dots, q_k$
*) Outputs q_i s.t. $q_i(D) - \mathcal{M}(X)_i = \max_{j=1}^{k} q_j(D) - \mathcal{M}(X)_j$

$\mathcal{M}$ is (ε, δ)-DP
$\Rightarrow \mathcal{W}$ is (ε, δ)-DP

post-processing

q_i has max error

Proof pt 2

$$\mathbb{E}\max_{i=1}^{k} q_i(D) - \mathcal{M}(X)_i = \mathbb{E}[q_i(D) - \mathcal{M}(X)_i \mid q_i = \omega(X)]$$

$$= \mathbb{E}[q_i(D) - q_i(X) \mid q_i = \omega(X)]$$

$$\leq e^{\varepsilon} - 1 + \delta \text{ by lemma}$$

$$+ \mathbb{E}[q_i(X) - \mathcal{M}(X)_i \mid q_i = \omega(X)]$$

$$\leq \mathbb{E}[\max_{j=1}^{k} q_j(X) - \mathcal{M}(X)_j]$$

$$\leq \alpha$$

$$\leq e^{\varepsilon} - 1 + \delta + \alpha.$$