Lecture 9
Sparse Vector Technique
Intro
$f_1, f_2, \ldots, f_k$ ← sens 1

Online

At time t, f_t arrives, must answer $f_t(D)$ privately

ε-DP. $f_t(D) + \text{Lap}(\frac{k}{\varepsilon})$

(ε,δ)-DP. $f_t(D) + \text{Lap}\left(\frac{\sqrt{k}}{\varepsilon}\right)$ ← Adv. comp

poly(k) error

Easier: which q's are "large"?

⇒ Offline: $\underbrace{f_1, \ldots, f_k}$. Exp mech. $\left.\begin{matrix} \\ \end{matrix}\right\}$ return largest query.

D dataset, objects, score $= f_i(D)$

Find c large q's, $\frac{c \log k}{\varepsilon}$ $\left(\frac{\sqrt{c} \log k}{\varepsilon}\right)$ approx

Online $(f_i(D) \geq T?)$

- Which q's are greater than T? ← public

  (Goal: output first c such q's).

AboveThreshold

Privacy

ε-DP

$D, D' \leftarrow$ databases

$A, A' \leftarrow$ outputs

$\perp \perp \perp \top$

$a = \perp^{t-1} \top$

$\Pr[A=a]$ vs $\Pr[A'=a]$ $\qquad \hat{T}, \nu_t$

Fix $\nu_1, \dots, \nu_{t-1}$

$g(D) = \max_{i \le t-1} (f_i(D) + \nu_i)$

$\uparrow$ deterministic

Algorithm 1 Input is a private database D, an adaptively chosen stream of sensitivity 1 queries $f_1, \dots$, and a threshold T. Output is a stream of responses $a_1, \dots$

AboveThreshold$(D, \{f_i\}, T, \epsilon)$
- **Let** $\hat{T} = T + \mathrm{Lap}\left(\frac{2}{\epsilon}\right)$.
- **for** Each query i **do**
  - **Let** $\nu_i = \mathrm{Lap}(\frac{4}{\epsilon})$
  - **if** $f_i(D) + \nu_i \ge \hat{T}$ **then**
    - **Output** $a_i = \top$.
    - **Halt.**
  - **else**
    - **Output** $a_i = \perp$.
  - **end if**
- **end for**

$$\Pr_{\hat{T}, \nu_t}[A=a] = \Pr[\hat{T} > g(D) \text{ and } f_t(D) + \nu_t \ge \hat{T}]$$

$$= \Pr[g(D) \le \hat{T} \le f_t(D) + \nu_t]$$

$$= \int\int_{-\infty}^{\infty} \Pr[\nu_t = v] \Pr[\hat{T} = \tau] \mathbb{1}\{\tau \in (g(D), f_t(D) + v)\} \, dv \, d\tau$$

$$\tau = \hat{\tau} - g(D') + g(D), \quad v = \hat{v} - g(D') + g(D) - f_t(D) + f_t(D')$$

$$= \int\int \Pr[\nu_t = v] \Pr[\hat{T} = \tau] \mathbb{1}\{(\hat{\tau} - g(D') + g(D)) \in (g(D), \cancel{f_t(D)} + \hat{v} - g(D') + g(D) - \cancel{f_t(D)} + f_t(D'))\}$$

$$= \int\int \Pr[\nu_t = v] \Pr[\hat{T} = \tau] \mathbb{1}\{\hat{\tau} \in (g(D'), \hat{v} + f_t(D'))\} \, d\hat{v} \, d\hat{\tau}$$

$$|\tau - \hat{\tau}| \le |g(D) - g(D')| \le 1. \quad |v - \hat{v}| \le 2.$$

$$\le \int\int \exp(\varepsilon/2) \Pr[\nu_t = \hat{v}] \exp(\varepsilon/2) \Pr[\hat{T} = \hat{\tau}] \mathbb{1}\{\hat{\tau} \in (g(D'), \hat{v} + f_t(D'))\} \, d\hat{v} \, d\hat{\tau}$$

$$\exp(\varepsilon) \cdot \Pr[\hat{T} \ge g(D') \text{ and } f_t(D') + \nu_t \ge \hat{T}]$$

$$= e^{\varepsilon} \Pr[A' = a].$$

Accuracy

"Mistakes"

(α, β) accurate: $\forall i \in [k]$, w.p. $1-\beta$

- If $a_i = \top$, $f_i(D) \geq T - \alpha$.
- If $a_i = \bot$, $f_i(D) \leq T + \alpha$.

Thm: k queries, $\forall i < k$, $f_i(D) \leq T - \alpha$.

Then. (α, β)-accurate for $\alpha = \underline{8(\log k + \log(2/\beta))}$ ε

Proof

We compare $f_i(D) + \overset{\mathrm{Lap}(\frac{4}{\varepsilon})}{\nu_i}$ vs $T + \overbrace{(\hat{T} - T)}^{\mathrm{Lap}(\frac{2}{\varepsilon})}$

Prove $\to$ $|\nu_i|$ (k rvs), $|\hat{T} - T| \leq \frac{\alpha}{2}$

$a_i = \top \Rightarrow f_i(D) + \nu_i \geq T + (\hat{T} - T)$

$f_i(D) \geq T + (\hat{T} - T) - \nu_i \geq T - |\alpha/2| - |\alpha/2| \geq T - \alpha$

$\Pr[|\mathrm{Lap}(b)| \geq tb] = \exp(-t)$

$\Pr[|\hat{T} - T| \geq \frac{\alpha}{2}] = \exp(-\varepsilon\alpha/4c) \leq \beta/2c$

$\Pr[\max_i |\nu_i| \geq \frac{\alpha}{2}] \leq \underline{k \cdot \exp(-\varepsilon\alpha/8c) \leq \beta/2c}$

$\log k - \frac{\varepsilon\alpha}{8} \leq \log(\beta/2)$

$\hookrightarrow \log k + \log(2/\beta) \leq \frac{\varepsilon\alpha}{8}$

$\alpha \geq \frac{8(\log k + \log(2/\beta))}{\varepsilon}$

Sparse Vector

c outputs

Thm: (ε,δ) DP, $\forall \varepsilon>0,\ \delta\geq 0$.

Algorithm 2 Input is a private database D, an adaptively chosen stream of sensitivity 1 queries $f_1, \ldots$, a threshold T, and a cutoff point c. Output is a stream of answers $a_1, \ldots$

Sparse$(D, \{f_i\}, T, c, \epsilon, \delta)$

- **If** $\delta = 0$ **Let** $\sigma = \frac{2c}{\epsilon}$. **Else Let** $\sigma = \frac{\sqrt{32c\ln\frac{1}{\delta}}}{\epsilon}$
- **Let** $\hat{T}_0 = T + \mathrm{Lap}(\sigma)$
- **Let** count $= 0$
- **for** Each query i **do**
  - **Let** $\nu_i = \mathrm{Lap}(2\sigma)$
  - **if** $f_i(D) + \nu_i \geq \hat{T}_{\text{count}}$ **then**
    - **Output** $a_i = \top$. $a_i = f_i(D) + \mathrm{Lap}(\theta(\frac{c}{\varepsilon}))$
    - **Let** count $=$ count $+1$.
    - **Let** $\hat{T}_{\text{count}} = T + \mathrm{Lap}(\sigma)$
  - **else**
    - **Output** $a_i = \bot$.
  - **end if**
  - **if** count $\geq c$ **then**
    - **Halt**.
  - **end if**
- **end for**

Union bd.

Theorem 4. *Suppose we are given a sequence of k queries where only c are large (i.e., the number of i such that $f_i(D) \geq T - \alpha$ is at most c). If $\delta = 0$, then Sparse is (α, β) accurate for $\alpha = \frac{8c(\log k + \log(2c/\beta))}{\varepsilon}$. If $\delta > 0$, then it is (α, β) accurate for $\alpha = \frac{\sqrt{512c\log(1/\delta)}(\log k + \log(2c/\beta))}{\varepsilon}$.*

Composition scaling ε.

NumericSparse

- Output values of large q's.

$\{\perp, \top\}^* \Longrightarrow (\mathbb{R} \cup \{\perp\})^*$

- If $a_i = \perp$, $f_i(D) \leq T + \alpha$
- If $a_i \in \mathbb{R}$, $|f_i(D) - a_i| \leq \alpha$.
  , $f_i(D) \geq T - \alpha$

$f(X) = \frac{1}{n}\sum f(X_i)$

$f: \mathcal{X}^n \to [0,1]$

Online Private Multiplicative Weights

$\hat{p}$ over $\mathcal{X}$

Sens-$1/n$

True dataset p

Q, $\forall f \in Q$, estimate $f(p)$ privately

Offline:

$\hat{p} = \mathrm{Unif}(\mathcal{X})$

1. Exp mech: choose $f \in Q$ w/ large error: $|f(\hat{p}) - f(p)|$ is large
2. Check how large $|f(\hat{p}) - f(p)|$ is. If small, return $\hat{p}$.
3. If large, MW update, $f(p)$ privately

$c = O\left(\frac{\log|\mathcal{X}|}{\alpha^2}\right)$

Online

k, $f_1, \dots, f_k \in Q$

Answer f_i at time i accurately

$\hat{p} = \mathrm{Unif}(\mathcal{X})$

Run sparse vector $|f_i(p) - f_i(\hat{p})|$, $T = \Theta(\alpha)$

1. If Sparse $\perp$: output $f_i(\hat{p})$
2. If sparse $\top$: output $f_i(p) + \mathrm{Lap}$
   ↳ Do an MW update using

Theorem 5. *The Online Private Multiplicative Weights algorithm can, in an online manner, answer a set Q of linear queries on a database of size n to accuracy α under (ε, δ)-DP, given $n = \tilde{O}\left(\frac{\log|Q|\sqrt{\log|\mathcal{X}|\log(1/\delta)}}{\alpha^2\varepsilon}\right)$ datapoints.*