<u>Sparse Vector / Above Threshold:</u>

Report Noisy Max: given k queries, output one with highest value.

- Intuition: since only outputting <u>name</u> of <u>one</u> query, need less noise than if actually answering all queries.

Today: related but slightly different setting.

- Given <u>stream</u> of queries $f_1, f_2, \ldots$, each has sensitivity 1.
- Public threshold T
- output <u>first</u> query above T, i.e., output
  $\min i : f_i(D) \geq T$.

- Can't use exponential mechanism, since online!
- Same w/ RNM but close
- Will want to generalize: output first c above T

- Turns out to be a super useful primitive, even in non-online settings (e.g. "online" b/c iterations of larger alg).

Idea: Almost same as RNM, but also use noisy threshold.

AboveThreshold:

- Let $\hat{T} = T + Lap(2/\varepsilon)$
- For each query i:
  - Let $\gamma_i = Lap(4/\varepsilon)$
  - if $f_i(D) + \gamma_i \geq \hat{T}$,
    - output i, halt

Thm: AboveThreshold is ε-DP

Pf: Consider some k, $D \sim D' \in \mathcal{X}$

Fix $\gamma_1, \dots, \gamma_{k-1}$. Take probs over $\gamma_k, \hat{T}$

$$g(D) = \max_{i<k} (f_i(D) + \gamma_i)$$

Abuse notation: $Pr[\hat{T} = t] = $ pdf of fn at t.

$$\Pr_{\hat{T}, \gamma_k}[A(D) = k] = Pr[\hat{T} \in (g(D), f_k(D) + \gamma_k]]$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Pr[\gamma_k = v] \cdot Pr[\hat{T} = t] \cdot \mathbb{1}[t \in (g(D), f_k(D) + \gamma_k]] \, dv \, dt$$

Change of vars: $\hat{v} = v + g(D) - g(D') + f_k(D') - f_k(D)$

$\hat{t} = t + g(D) - g(D')$

Note: $|\hat{v} - v| \leq 2$, $|\hat{t} - t| \leq 1$, since

$g(D) - g(D') \leq 1$, $f_k(D') - f_k(D) \leq 1$

$$\hat{T} = t + g(D) - g(D')$$
$$\gamma_k = v + g(D) - g(D') + f_k(D') - f_k(D)$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\gamma_k = \hat{v}] \cdot \Pr[\hat{T} = \hat{t}] \cdot \mathbb{1}[\hat{t} \in (g(D), f_k(D) + \hat{v}]] \, dv \, dt$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\gamma_k = \hat{v}] \cdot \Pr[\hat{T} = \hat{t}] \cdot \mathbb{1}[t + g(D) - g(D') \in (g(D), v + g(D) - g(D') + f_k(D')]] \, dv \, dt$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \Pr[\gamma_k = \hat{v}] \cdot \Pr[\hat{T} = \hat{t}] \cdot \mathbb{1}[t \in (g(D'), f_k(D') + v]] \, dv \, dt$$

$\mathrm{Lap}(4/\varepsilon)$, sensitivity 2 $\qquad$ $\mathrm{Lap}(2/\varepsilon)$, sensitivity 1

$$\leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(\varepsilon/2) \Pr[\gamma_k = v] \cdot \exp(\varepsilon/2) \Pr[\hat{T} = t] \cdot \mathbb{1}[t \in (g(D'), f_k(D') + v]] \, dv \, dt$$

$$= \exp(\varepsilon) \Pr_{\gamma_k, \hat{T}}[\hat{T} \in (g(D'), f_k(D') + \gamma_k]]$$

$$= \exp(\varepsilon) \cdot \Pr_{\hat{T}, \gamma_k}[A(D') = k]$$

Accuracy:

Def: (α,β)-accurate if with prob $\geq 1-\beta$:

- Any k output by alg has $f_k(D) \geq T-\alpha$
- Any i not output by alg has $f_k(D) \leq T+\alpha$

Let $\beta \in (0,1)$, and let $\alpha = \frac{8(\log k + \log \frac{2}{\beta})}{\varepsilon}$

Thm: If $f_i(D) < T-\alpha$ $\forall i < k$, then Above Threshold is (α,β)-accurate.

Pf: Suff. $\max_{i \in [k]} |\gamma_i| + |T-\hat{T}| \leq \alpha$

Then if we output i:

$f_i(D) + \gamma_i \geq \hat{T} \Rightarrow f_i(D) \geq T - |T-\hat{T}| - |\gamma_i|$

$\geq T-\alpha$ ✓

If we don't output i:

$f_i(D) + \gamma_i < \hat{T} \leq T + |T-\hat{T}|$

$\Rightarrow f_i(D) < T + |T-\hat{T}| + |\gamma_i| \leq T+\alpha$ ✓

So just WTS that $\max_{i \in [k]} |\gamma_i| + |T-\hat{T}| \leq \alpha$ w.p. $\geq 1-\beta$

Laplace tail bound: If $Y \sim Lap(b)$, $Pr[|Y| \geq t \cdot b] = e^{-t}$

$$\Rightarrow \Pr[|T-\hat{T}| \geq \tfrac{\alpha}{2}] = \exp\left(-\tfrac{\varepsilon\alpha}{4}\right) = \exp\left(-2\left(\log k + \log \tfrac{2}{\beta}\right)\right)$$

$$\leq \exp\left(-\log \tfrac{2}{\beta}\right) = \tfrac{\beta}{2}$$

Similarly, apply union bound to γ_i's:

$$\Pr\left[\max_{i \in [k]} |\gamma_i| \geq \tfrac{\alpha}{2}\right] \leq k \cdot \exp\left(-\tfrac{\varepsilon\alpha}{8}\right) = k \cdot \exp\left(-\left(\log k + \log \tfrac{2}{\beta}\right)\right)$$

$$= k \cdot \tfrac{1}{k} \cdot \tfrac{\beta}{2} = \tfrac{\beta}{2} \quad \checkmark$$

Union bound ✓

Now generalize: want to output first c queries above T.

Idea: just compose c runs of Above threshold!

Sparse:

- If $\delta = 0$ let $\sigma = \frac{2c}{\varepsilon}$. Else let $\sigma = \frac{\sqrt{32 c \ln \frac{1}{\delta}}}{\varepsilon}$
- Let $\hat{T}_0 = T + \mathrm{Lap}(\sigma)$
- Let count $= 0$
- For each query i:
  - Let $\gamma_i = \mathrm{Lap}(2\sigma)$
  - if $f_i(D) + \gamma_i \geq \hat{T}_{\text{count}}$:
    - output i

- count ++
- Let $\hat{T}_{count} \sim T + Lap(\sigma)$
- if count $\geq c$ Halt.

<u>Thm</u>: Sparse is (ε, δ)-DP

<u>Pf</u>: Sparse equivalent to running Above Threshold

w/ $\varepsilon' = \begin{cases} \frac{\varepsilon}{c} & \text{if } \delta = 0, \\ \frac{\varepsilon}{\sqrt{8c \ln \frac{1}{\delta}}} & \text{if } \delta > 0, \end{cases}$

restarting w/ fresh randomness on each output, $\leq c$ times

If $\delta = 0$: Basic composition $\Rightarrow \leq c \cdot \varepsilon' = \varepsilon$-DP ✓

If $\delta > 0$: Advanced composition: $\sqrt{2k \ln \frac{1}{\delta}} \cdot \varepsilon' + k\varepsilon'(e^{\varepsilon'} - 1)$

$$\varepsilon' \to \frac{\varepsilon}{\sqrt{8c \ln \frac{1}{\delta}}} \cdot \sqrt{2c \ln \frac{1}{\delta}} + c \frac{\varepsilon^2}{8c \ln \frac{1}{\delta}}$$

$$= \frac{\varepsilon}{2} + \frac{\varepsilon^2}{8 \ln \frac{1}{\delta}} \leq \varepsilon \qquad \left(\varepsilon \leq 4 \ln \frac{1}{\delta}\right)$$

Accuracy:

<u>Idea</u>: If each call to Above Threshold is $(\alpha, \beta/c)$-accurate, Sparse is (α, β)-accurate. (union bound)

Thm: Suppose $L(T) = |\{i \le k : f_i(D) \ge T - \alpha\}| \le c$.

If $\delta > 0$, Sparse is (α, β)-accurate for

$$\alpha = \frac{(\ln k + \ln \frac{2c}{\beta})\sqrt{512 c \ln \frac{1}{\delta}}}{\varepsilon}$$

If $\delta = 0$, Sparse is (α, β)-accurate for

$$\alpha = \frac{8c(\ln k + \ln \frac{2c}{\beta})}{\varepsilon}$$

Pf: Plug in accuracy for AT with

$\varepsilon = \varepsilon'$, $\beta = \beta/c$.

Union bound.

Numeric Sparse: can also output values for queries above threshold!

- Use Laplace mechanism for each, only doubles privacy loss via basic composition.

Accuracy: see book. Informally, (α, β)-accurate if w.p. $\ge 1 - \beta$, any value output within α of truth, any not output has $f_i(D) \le T + \alpha$

Punchline: total privacy loss similar to if we
<u>knew</u> which queries were above T!
- Finding big queries is free!