Linear Queries w/ correlated error:

- Sps class of queries Q (known), each sensitivity 1.
- Can use Laplace mechanism, composition
  - basic: noise scales like $|Q|$
  - advanced: noise scales like $\sqrt{|Q| \ln 1/\delta}$

Basically tight: Q could just be many copies of same query, need to increase noise as ask more queries!

But then trivial: return same answer each time!

In general: if Q "nice", e.g., not many distinct queries, can we do better?

Goal: noise grows w/ $\log |Q|$, or better.

Offline: Q known. Technique: release synthetic database that's correct on queries in Q. Exponential mechanism.

Online: Q not known ahead of time. Can still get similar bounds. "Private Multiplicative Weights", sparse vector technique.

Today: offline

Setup: Linear queries.

- Generalize counting queries
- Let X = domain of databases, so $P(X) = \mathcal{D}$ (with multiplicities)

think: every possible row of database

- Given $D \in \mathcal{D}$, $x \in X$, let D_x = # copies of x in D

$q: X \to \{0,1\}$ predicate: "does this row correspond to having cancer?"

- Counting query: $f_q(D) = \sum_{x \in X} D_x \, q(x)$
- Normalized counting query: $f_q(D) = \frac{1}{|D|} \sum_{x \in X} D_x \, q(x)$

- Linear queries: $q: X \to [0,1]$

$$f_q(D) = \sum_{x \in X} D_x \, q(x), \text{ or } f_q(D) = \frac{1}{|D|} \sum_{x \in X} D_x \, q(x)$$

Note: $\Delta f_q \le 1$ unnormalized, $\Delta f_q \le \frac{1}{|D|}$ normalized

Ex: Marginal Tables

Say $X = \{0,1\}^d$, collection of boolean features (college grad, US citizen, family history of cancer...)

Queries: what fraction of the dataset have features a, b, c?

$S \subseteq d$, $q_S(x) = \prod_{i \in S} x_i$

"Marginals"

$Q = \{ f_{q_S} : S \subseteq \{1, 2, \dots d\} \}$ all marginals

$|Q| = 2^d$

$Q = \{ f_{q_S} : S \subseteq [d], |S| \leq k \}$ k-way marginals

$|Q| = \binom{d}{k} \approx d^k$

<u>Lots</u> of marginals, so really want to add noise $\sim \log |Q|$ rather than $\sqrt{|Q|}$ or $|Q|$!

<u>Offline</u>: SmallDB$(D, Q, \varepsilon, \alpha)$

- Let $R = \{ D \in \mathcal{D} : |D| = \frac{\log |Q|}{\alpha^2} \}$ $\leftarrow$ small!
- Let $u : \mathcal{D} \times R \to \mathbb{R}$ be

$$u(D, \hat{D}) = - \max_{f \in Q} |f(D) - f(\hat{D})|$$

- Use exponential mechanism $M_E(D, u, R)$ to sample small database from R

<u>Thm</u>: ε-DP

<u>Pf</u>: Just exponential mechanism!

WTS: There is a small database that's good.

Lemma: Let $D \in \mathcal{D}$, Q collection of (normalized) linear queries. There exists $\hat{D}$ with $|\hat{D}| = \frac{\log|Q|}{\alpha^2}$ s.t.

$$\max_{f \in Q} |f(D) - f(\hat{D})| \leq \alpha$$

Pf: Let $m = \frac{\log|Q|}{\alpha^2}$.

Construct $\hat{D}$ by sampling m entries from D uniformly at random.

Let Y_i be i'th sample

Let $f_q \in Q$.

$$\Rightarrow f_q(\hat{D}) = \frac{1}{n} \sum_{x \in X} \hat{D}_x \, q(x) = \frac{1}{m} \sum_{i=1}^{m} q(Y_i)$$

Note: $0 \leq q(Y_i) \leq 1$, and

$$E[q(Y_i)] = \sum_{x \in D} \frac{1}{|D|} q(x) = f_q(D)$$

$$\Rightarrow E[f_q(\hat{D})] = \frac{1}{m} \sum_{i=1}^{m} E[q(Y_i)] = \frac{1}{n} \sum_{i=1}^{n} f_q(D) = f_q(D)$$

Hoeffding bound (additive Chernoff):

Let $X_1, \dots, X_m$ independent random vars s.t. $0 \leq X_i \leq 1$ $\forall i$.

Then $\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i > E\left[\frac{1}{m}\sum_{i=1}^{m} X_i\right] + \varepsilon\right] \le \exp(-2m\varepsilon^2)$

$$< \qquad\qquad - \varepsilon\Big] \le \exp(-2m\varepsilon^2)$$

$$\underbrace{E[f_q(\hat{D})]}$$

$$\Rightarrow \Pr\left[|f_q(\hat{D}) - f_q(D)| > \alpha\right] \le 2e^{-2m\alpha^2}$$

Union bound over all $f \in Q$:

$$\Pr\left[\max_{f \in Q} |f(D) - f(\hat{D})| > \alpha\right] \le 2|Q| e^{-2m\alpha^2}$$

$$m = \frac{\log|Q|}{\alpha^2} \quad \Rightarrow \quad 2|Q|e^{-2m\alpha^2} = 2|Q|e^{-2\log|Q|}$$

$$= \frac{2|Q|}{|Q|^2} < 1 \quad (|Q| > 2)$$

$\Rightarrow \exists$ good database of size m ✓

So a good database exists. But we run exponential mechanism

Lemma: With prob. $\ge 1-\beta$,

$$\max_{f \in Q} |f(D) - f(\hat{D})| \le \alpha + \frac{2\left(\frac{\log|X| \cdot \log|Q|}{\alpha^2} + \log\frac{1}{\beta}\right)}{\varepsilon|D|}$$

Pf: use utility bound for exponential mechanism:

$$\Pr[u(M_\varepsilon(D,u,R)) \le OPT_u(D) - \frac{2\Delta u}{\varepsilon}(\log|R| + t)] \le e^{-t}$$

$\ge -\alpha$ by lemma $\qquad \le \frac{1}{|D|}$

$$\Rightarrow \Pr[\max_{f\in Q}|f(D)-\hat{f}(D)| \ge -OPT_u(D) + \frac{2\Delta u}{\varepsilon}(\log|R| + t)] \le e^{-t}$$

$$\Rightarrow \Pr[\max_{f\in Q}|f(D)-\hat{f}(D)| \ge \alpha + \frac{2}{\varepsilon|D|}(\log(|X|^{\frac{\log|Q|}{\alpha^2}}) + \ln\frac{1}{\beta})] \le \beta$$

$$\Rightarrow \Pr[\max_{f\in Q}|f(D)-f(\hat{D})| \ge \alpha + \frac{2}{\varepsilon|D|}(\frac{\log|Q|\log|X|}{\alpha^2} + \ln\frac{1}{\beta})] \le \beta \checkmark$$

Thm: For any database D with

$$|D| \ge \frac{16\log|X|\log|Q| + 4\alpha^2\log\frac{1}{\beta}}{\varepsilon\alpha^3}$$

w.p. $\ge 1-\beta$, $\max_{f\in Q}|f(D)-f(\hat{D})| \le \alpha$

Pf: use previous lemma w/ $\frac{\alpha}{2}$: w.p. $\ge 1-\beta$

$$\max_{f\in Q}|f(D)-f(\hat{D})| \le \frac{\alpha}{2} + \frac{2(\frac{4\log|X|\log|Q|}{\alpha^2} + \log\frac{1}{\beta})}{\varepsilon|D|}$$

set to be $\le\alpha$, solve for $|D|$:

$$\frac{\alpha}{2} + \frac{2(\frac{4\log|X|\log|Q|}{\alpha^2} + \log\frac{1}{\beta})}{\varepsilon|D|} \le \alpha$$

$$\Leftrightarrow \frac{\alpha}{2}|D| + \frac{1}{\varepsilon}\left(\frac{8\log|X|\log|Q|}{\alpha^2} \; 2\log\frac{1}{\beta}\right) \le \alpha|D|$$

$$\Leftrightarrow |D| \ge \frac{1}{\varepsilon}\left(\frac{16\log|X|\log|Q| + 4\alpha^2\log\frac{1}{\beta}}{\alpha^3}\right)$$

interpretation: Think of α small constant, e.g., $\frac{1}{10}$

Can answer all queries w/ database of size

$\simeq \frac{1}{\varepsilon}\log|X|\log|Q|$!

e.g. all marginals w/ $\approx \frac{1}{\varepsilon} d^2$!

More refined bounds:

VC-dimension of Q replaces $\log|Q|$ for counting queries

fat-shattering dim of Q for linear queries.