Setup: Linear queries.

- Generalize counting queries
- Let X = domain of databases, so $\mathcal{P}(X) = \mathcal{D}$ (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: $\Delta f_q \leq 1$ unnormalized, $\Delta f_q \leq \frac{1}{|D|}$ normalized

Offline: SmallDB$(D, Q, \varepsilon, \alpha)$

- Let $R = \left\{ D \in \mathcal{D} : |D| = \frac{\log |Q|}{\alpha^2} \right\}$ ← 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

Thm: ε-DP

Pf: Just exponential mechanism!

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.

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

So, $\exists$ good small database. But we run exponential mechanism. Is it good?

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:

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

$OPT_u(D) \ge -\alpha$ by lemma; $\Delta u \le \frac{1}{|D|}$

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

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

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

<u>Thm</u>: For any database D with

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

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

<u>Pf</u>: Use previous lemma w/ $\frac{\alpha}{2}$: w.p. $\geq 1-\beta$

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

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

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

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

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

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

Can answer all queries w/ database of size

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

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

Thm: For any $D \in \mathcal{X}$, w.p. $\geq 1-\beta$,

$$\max_{f \in Q} |f(D) - f(\hat{D})| \leq \left(\frac{16 \log|X| \log|Q| + 4 \log \frac{1}{\beta}}{\varepsilon |D|} \right)^{1/3}$$

Pf: Same, solve for α

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

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

$$\Rightarrow \alpha^3 \geq \frac{16 \log|X| \log|Q| + 4 \log \frac{1}{\beta}}{\varepsilon |D|} \quad \leftarrow \alpha \leq 1$$

$$\Rightarrow \alpha \geq \left(\frac{16 \log|X| \log|Q| + 4 \log \frac{1}{\beta}}{\varepsilon |D|} \right)^{1/3}$$

More refined bounds:

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

fat-shattering dim of Q for linear queries.

Online Setting:

- Don't know Q, get (linear) queries online.
- "private multiplicative weights":
  combination of MWU (Hedge) and (Numeric) Sparse Vector
- MWU: super useful meta-algorithm!
  - Online learning (no-regret algorithms), AdaBoost
  - Fast flow algorithms
  - Fast LP solving

    ⋮

Super high level idea: (no privacy)

- maintain synthetic database $\hat{D}$
- On query f^t:
  - If $|f^t(D) - f^t(\hat{D})|$ small, output $f^t(\hat{D})$ ← Sparse vector!
  - Else update $\hat{D}$ to have smaller error on f^t, output $f^t(D) +$ noise
    ↑ MWU   ↑ (Laplace).

Since using sparse vector, only (really) pay when error large.

MWU ⟹ can't be large too many times!

<u>First</u>: MWU (Hedge) in this language, convergence guarantee.

<u>Second</u>: Private Version

<u>Recall</u>: D_i = # copies of i in D, where $i \in X$, $D \in \mathcal{D}$.

MWU usually thought of as maintaining probability distribution, so rescale: divide all D_i by $|D|$

so $\sum_{i \in X} D_i = 1$.

⟹ doesn't affect normalized linear queries!

- Notation: $f \in Q$; $\sum_{i \in X} f_i D_i$ $\quad (f_i \in [0,1])$

current synthetic database

current query

"true" answer.

$MW(D^t, f^t, v^t)$

- if $v^t < f^t(D^t)$

  - Let $r^t = f^t$

- else $(v^t \geq f^t(D^t))$

  - Let $r^t = 1 - f^t$ $\quad$ (i.e., $r_i^t = 1 - f_i^t \;\; \forall i \in X$)

- Update: $\forall i \in X$, let

- $\hat{D}_i^{t+1} = \exp\left(-\frac{\alpha}{2} r_i^t\right) D_i^t$

- $D_i^{t+1} = \dfrac{\hat{D}_i^{t+1}}{\sum_{i \in X} \hat{D}_i^{t+1}}$

- Output D^{t+1}

Thm: Let Q set of linear queries, $D \in \mathcal{D}$ (normalized so $\sum_{i \in X} D_i = 1$), and let $D_i^1 = \frac{1}{|X|}$ $\forall i \in X$.

Let D^t for $t \in \{2, \dots, L\}$ generated by setting $D^t = MW(D^{t-1}, f^t, v^t)$, where for each t, $f^t \in Q$, and

1) $|f^t(D) - f^t(D^t)| > \alpha$, and

2) $|f^t(D) - v^t| < \alpha$.

Then $L \le 1 + \dfrac{4 \log |X|}{\alpha^2}$

Cor: Once we've run MW $1 + \frac{4 \log |X|}{\alpha^2}$ times, final database must have error $\le \alpha$ on all queries in Q, or else could extend sequence!

$\Rightarrow$ never need to update again, no further privacy loss.

Future: alg will use sparse vector to test if $|f(D) - f(D')| > \alpha$ on query f,

If so, will call $MW(D^t, f, f(D) + Lap)$ to update

$\Rightarrow$ at most $1 + \frac{4 \log |X|}{\alpha^2}$ updates!

Pf: Potential fn Ψ keeping track of (dis)similarity of D^t and D. Show

1) Ψ not too large at beginning
2) Goes down a lot in each round
3) Always nonnegative

Use KL-divergence!

$$\Psi_t = KL(D \| D^t) = \sum_{i \in X} D_i \log\left(\frac{D_i}{D^t_i}\right)$$

Lemma: $\Psi_t \geq 0 \quad \forall t$

Pf: log-sum inequality.

If $a_1, \ldots, a_n \geq 0$, $b_1, \ldots, b_n \geq 0$, then

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i\right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

$$\Rightarrow \Psi_t \geq 1 \log \frac{1}{1} = 0$$

Lemma: $\Psi_1 \leq \log |X|$

Pf: $\Psi_1 = \sum_{i \in X} D_i \log(|X| D_i) \leq \log n$ (D = prob. dist.)

Lemma: $\Psi_t - \Psi_{t+1} \geq \frac{\alpha}{2}(r^t(D^t) - r^t(D)) - \frac{\alpha^2}{4}$

Pf: $\Psi_t - \Psi_{t+1} = \sum_{i \in X} D_i \log\left(\frac{D_i}{D_i^t}\right) - \sum_{i \in X} D_i \log\left(\frac{D_i}{D_i^{t+1}}\right)$

$$= \sum_{i \in X} D_i \log\left(\frac{D_i^{t+1}}{D_i^t}\right)$$

$$= \sum_{i \in X} D_i \log\left(\frac{\frac{\hat{D}_i^{t+1}}{\sum_{j \in X} \hat{D}_j^{t+1}}}{D_i^t}\right)$$

$$= \sum_{i \in X} D_i \left(\log\left(\frac{\hat{D}_i^{t+1}}{D_i^t}\right) - \log\left(\sum_{j \in X} \hat{D}_j^{t+1}\right)\right)$$

$$= \sum_{i \in X} D_i \left(\log\left(\frac{D_i^t \cdot \exp(-\frac{\alpha}{2} r_i^t)}{D_i^t}\right) - \log\left(\sum_{j \in X} D_j^t \exp(-\frac{\alpha}{2} r_j^t)\right)\right)$$

$$= -\left(\sum_{i \in X} D_i \frac{\alpha}{2} r_i^t\right) - \log\left(\sum_{j \in X} D_j^t \exp(-\frac{\alpha}{2} r_j^t)\right)$$

$$= -\frac{\alpha}{2} r^t(D) - \log \sum_{j \in X} D_j^t \exp(-\frac{\alpha}{2} r_j^t)$$

useful fact: $e^{-x} \leq 1 - x + x^2$ (for $x < 1.79$)

$$\geq -\frac{\alpha}{2} r^t(D) - \log \sum_{j \in X} D_j^t \left(1 - \frac{\alpha}{2} r_j^t + \frac{\alpha^2}{4} (r_j^t)^2\right)$$

$$\geq -\frac{\alpha}{2} r^t(D) - \log \sum_{j \in X} D_j^t \left(1 - \frac{\alpha}{2} r_j^t + \frac{\alpha^2}{4}\right)$$

$$= -\frac{\alpha}{2} r^t(D) - \log\left(1 + \frac{\alpha^2}{4} - \frac{\alpha}{2} r^t(D^t)\right)$$

useful fact: $\log(1+y) \leq y$ for $y > -1$

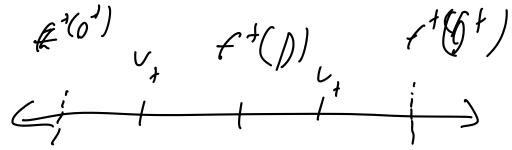
$$\geq -\frac{\alpha}{2} r^t(D) - \left(\frac{\alpha^2}{4} - \frac{\alpha}{2} r^t(D^t)\right)$$

$$= \frac{\alpha}{2}\left(r^t(D^t) - r^t(D)\right) - \frac{\alpha^2}{4} \quad \checkmark$$

we know that $\forall t$,

1) $|f^t(D) - f^t(D^t)| > \alpha$, and

2) $|f^t(D) - v^t| < \alpha$.



So $f^t(D) \le f^t(D^t) \iff v^t < f^t(D^t)$

If $r^t = f^t \Rightarrow v^t < f^t(D^t) \Rightarrow f^t(D) \le f^t(D^t)$

$$\Rightarrow r^t(D^t) - r^t(D) = f^t(D^t) - f^t(D) > \alpha$$

If $r^t = 1 - f^t \Rightarrow v^t \ge f^t(D^t) \Rightarrow f^t(D) > f^t(D^t)$

$$\Rightarrow r^t(D^t) - r^t(D) = \sum_{i \in X} r_i^t D_i^t - \sum_{i \in X} r_i^t D_i$$

$$= \sum_{i \in X} r_i^t (D_i^t - D_i)$$

$$= \sum_{i \in X} (1 - f_i^t)(D_i^t - D_i)$$

$$= \sum_{i \in X} D_i^t - \sum_{i \in X} D_i - \sum_{i \in X} f_i^t D_i^t + \sum_{i \in X} f_i^t D_i$$

$$= f^t(D) - f^t(D^t)$$

$$> \alpha$$

Plug into previous lemma:

$$\Psi_t - \Psi_{t+1} \ge \frac{\alpha}{2}\left(r^t(D^t) - r^t(D)\right) - \frac{\alpha^2}{4}$$

$$\geq \frac{\alpha}{2} \cdot \alpha - \frac{\alpha^2}{4}$$

$$= \frac{\alpha^2}{4}$$

So KL divergence goes down by $\geq \frac{\alpha^2}{4}$ each time

$$\Rightarrow 0 \underset{\text{nonnegative}}{\leq} \Psi_L \leq \Psi_1 - L \cdot \frac{\alpha^2}{4} \underset{\text{prev. lemma}}{\leq} \log |X| - L \cdot \frac{\alpha^2}{4}$$

$$\Rightarrow L \leq \frac{4 \log |X|}{\alpha^2}$$

✓