2/20/25: Finishing up private MWU for online linear queries

$MW(D^t, f^t, v^t)$ ← Note: doesn't access D! ⇒ post-processing

- if $v^t < f^t(D^t)$
  - Let $r^t = f^t$
- else $(v^t \geq f^t(D^t))$
  - Let $r^t = 1 - f^t$ (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}

<u>Thm</u>: 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 $t \in \{2, \ldots, L\}$ 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 $L \leq 1 + \dfrac{4 \log |X|}{\alpha^2}$

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 $a_1, \dots, a_n \geq 0$ $b_1, \dots, b_n \geq 0$, 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: $\Psi_t - \Psi_{t+1} \geq \frac{\alpha}{2}\left(r^t(D^t) - r^t(D)\right) - \frac{\alpha^2}{4}$

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

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

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

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

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

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

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

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 $y > 1$

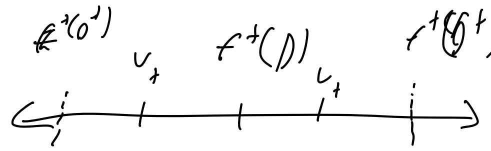
$$\geq -\frac{\alpha}{2} r^t(D) - \left(\frac{\alpha^2}{4} - \frac{\alpha}{2} r^t(D^t)\right)$$

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

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 $2\frac{\alpha^2}{4}$ each time

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

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

Next: combine MW with (numerical sparse vector (NumericSparse).

Private MW:

- Let $c = \frac{4 \log |X|}{\alpha^2}$
- Set T (threshold for sparse vector):

$$T = \frac{(2 + 32\sqrt{2}) \sqrt{c \log \frac{2}{\delta}} \left(\log k + \log \frac{4c}{\beta}\right)}{\varepsilon |D|}$$ (diff if $\delta = 0$)

- Init NumericSparse w/ threshold T, ε, δ, D, #above $= c$
- Let $t = 0$; let $D_i^0 = \frac{1}{|X|}$ $\forall i \in X$

- On linear query f_i:
  - Let $f'_{2i-1}(\cdot) = f_i() - f_i(D^t)$
  - Let $f'_{2i}() = f_i(D^t) - f_i()$
  - Let $E_{2i-1} = \text{NumericSparse}(f'_{2i-1})$
    $E_{2i} = \text{NumericSparse}(f'_{2i})$
  - If $E_{2i-1} = E_{2i} = \perp$ (below threshold)
    - $a_i = f_i(D^t)$
  - Else
    - If $E_{2i-1} \in \mathbb{R}$ (above threshold).
      - $a_i = f_i(D^t) + E_{2i-1}$
    - Else ($E_{2i} \in \mathbb{R}$)
      - $a_i = f_i(D^t) - E_{2i}$
    - $D^{t+1} = MW(D^t, f_i, a_i)$
    - $t{+}{+}$

Thm: Alg is (ε, δ)-DP

Pf: Only accesses DB through NumericSparse

$\Rightarrow$ (ε, δ)-DP since NumericSparse (ε, δ)-DP.

Utility:

Intuition: utility thm for NumericSparse

$\Rightarrow$ w.h.p., MW only invoked when

1) $|f_i(D) - f_i(D^t)|$ large

2) $|a_i - f_i(D)|$ small

$\Rightarrow$ MW convergence thm $\Rightarrow$ total # updates $\leq c$

$\Rightarrow$ correctly answer all queries.

Thm: With prob. $\geq 1-\beta$, $\forall$ queries f_i, Private MW returns an answer s.t. $|f_i(D) - a_i| \leq 3\alpha$

for any α s.t.

$$\alpha > \frac{(2+32\sqrt{2})\sqrt{\log|X| \log \frac{2}{\delta}}\left(\log|Q| + \log\left(\frac{32\log|X|}{\alpha^2\beta}\right)\right)}{\alpha \varepsilon |D|}$$

Optimize for α and un-normalize (so queries in $[0,|D|]$ rather than $[0,1]$)

$$\Rightarrow \text{error} = |D|^{1/2}\left(\frac{1}{\varepsilon} \cdot \Theta\left(\sqrt{\log|X| \log \frac{2}{\delta}}\left(\log|Q| + \log \frac{|D|}{\beta}\right)\right)\right)^{1/2}$$

Sparse DB: similar, but $|D|^{2/3}$!

PF: Numeric Sparse guarantee:

k queries, max c above threshold,

(α, β)-accurate for $\alpha \geq \frac{(\ln k + \ln \frac{4c}{\beta}) \sqrt{c \ln \frac{2}{\delta}} (\sqrt{512}+1)}{\varepsilon}$

For us: $k = 2|Q|$, and normalizing so everything divided by $|D|$

$c = \frac{4 \log |X|}{\alpha^2}$

$\Rightarrow (\alpha, \beta)$-accurate for $\alpha = 2T$

Remember (α, β)-accurate: w.p. $\geq 1-\beta$, 1) all times output a $\# z_i$, $|f_i(D) - z_i| \leq \alpha$ and $f_i(D)| \geq T - \alpha$

2) all times output $\perp$, $f_i(D) \leq T + \alpha$

So assume in the $1-\beta$ prob. case.

$\Rightarrow \forall i$ s.t. MW triggered,

$|f_i(D) - f_i(D^+)| \geq T - \alpha = 2\alpha - \alpha = \alpha$ and

$|a_i - f_i(D)| = |f_i(D) - f_i(D^+)| - E$

$\leq \alpha$

$\Rightarrow$ can apply convergence thm for MW!

$\leq c$ updates

Moreover:

- Can't have both f_{2i-1} and $f_{2i} \in \mathbb{R}$:

would imply that

$$f_i(D) - f_i(D') \geq T - \alpha \geq \alpha$$

$$f_i(D') - f_i(D) \geq T - \alpha \geq \alpha$$

$\cancel{\alpha}$

- If no update triggered $(a_i = f_i(D'))$, know

$$|f_i(D) - f_i(D')| \leq T + \alpha = 3\alpha$$

- If update triggered, $|a_i - f_i(D)| \leq \alpha$