

Theory of Differential Privacy

Lecture 7:

Sparse Vector Technique

EN.601.438/638 Fall 2026

Lydia Zakynthinou

Johns Hopkins University

Sept 23, 2026

Setup

- Queries (public)

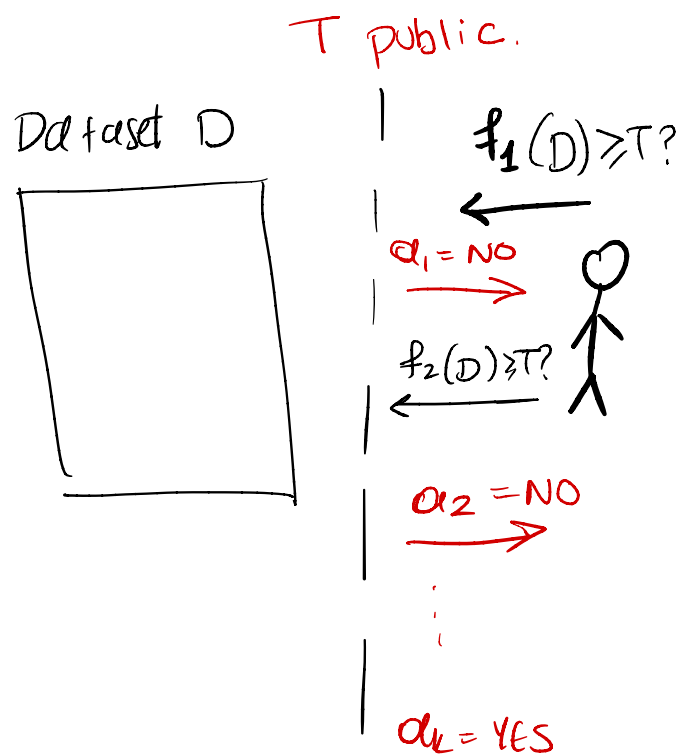
$$f_i: \mathcal{X}^n \rightarrow \mathbb{R}$$

$$\Delta f_i = \max_{x \sim x'} |f_i(x) - f_i(x')| \leq 1.$$

They arrive sequentially and may be chosen adaptively.

- Public threshold T .

- Output: stream α s.t. $\alpha_i \in \{NO, YES\}$ (YES means "the query is approx. above T ").



Above Threshold ($x, T, \epsilon, \{f_i\}$).

1. Draw $\hat{T} = T + \eta$ where $\eta \sim \text{Lap}(\frac{2}{\epsilon})$.

2. For $i=1, 2, \dots$

$$v_i \sim \text{Lap}(\frac{4}{\epsilon}).$$

If $f_i(x) + v_i \geq \hat{T}$, return $a_i = \text{YES}$ and stop.

Otherwise return $a_i = \text{NO}$ and continue.

Privacy Above Threshold is ϵ -DP.

← Does not require max # iterations to be bounded!

Proof sketch:

Fix neighboring $x \sim x'$. Fix the analyst's random coins and a transcript

$$\alpha = (\underbrace{\text{NO}, \text{NO}, \text{NO}, \dots}_{k-1}, \text{YES})$$

Then the analyst submits the same functions $f_1, \dots, f_k$.

Condition on arbitrary values $v_1, \dots, v_{k-1}$.

Define

$$g_x = \max_{i \leq k} (f_i(x) + v_i) \quad , \quad g_{x'} = \max_{i \leq k} (f_i(x') + v_i)$$

Transcript α occurs: when
(under input x)

$$g_x < \hat{T} \leq \max_k (f_k(x) + v_k)$$

Note: sensitivity of g_x is $|g_x - g_{x'}| \leq 1$.

Let $P_b(z) = \frac{1}{2b} e^{-|z|/b}$ denote the density of $\text{Lap}(b)$.

Let $P_T(t) = \text{Lap}(T, \frac{2}{\varepsilon}) = P_{2/\varepsilon}(t-T)$ be the density of $\hat{T}$.

Let $P_Q(v) = P_{4/\varepsilon}(v)$ be the density of v_k

Then:

$$\begin{aligned} P_{\mathcal{V}}[A(x) = \alpha / [v_i]_{i=1}^k] &= P_{\eta, v_k} [g_x < \hat{T} \leq f_k(x) + v_k] \\ &= \iint P_T(t) \cdot P_Q(v) \cdot \mathbb{1}_{\{g_x < t < f_k(x) + v\}} dt dv. \quad (1) \end{aligned}$$

Define $\Delta g = g_{x'} - g_x$, $\Delta f = f_k(x) - f_k(x')$.

Note: $|\Delta g| \leq 1$, $|\Delta f| \leq 1$

Note change of variables $\boxed{t' = t + \Delta g, \quad v' = v + \Delta f + \Delta g} \quad (2)$

- $g_x < t \iff g_{x'} < t'$
- $t \leq f_k(x) + v \iff t' \leq f_k(x') + v'$

I want to relate (1) to the same integral with $g_{x'}, f_k(x')$ instead

Proof.

$$f_k(x') + v' = f_k(x') + v + \Delta f + \Delta g = \cancel{f_k(x')} + v + f_k(x) - \cancel{f_k(x)} + \Delta g$$

$$= f_k(x) + v + \Delta g \quad (*)$$

$$f_k(x') + v' \geq t' \stackrel{(*)}{\Leftrightarrow} f_k(x) + v + \Delta g \geq \overbrace{t'}^{t' + \Delta g} \Leftrightarrow f_k(x) + v \geq t. \quad (\checkmark)$$

Substitute (2) in (1):

$$(1) = \iint P_T(t' - \Delta g) \cdot P_Q(v' - \Delta g - \Delta f) \cdot \mathbb{1}\{g_{x'} < t' \leq f_k(x') + v'\} dt' dv'.$$

fact. $P_b(z-s) \leq e^{|s|/b} \cdot P_b(z)$

- $P_T(t' - \Delta g) = P_{2/\varepsilon}(t' - \Delta g) \leq e^{|\Delta g|/(2/\varepsilon)} \cdot P_{2/\varepsilon}(t') \stackrel{|\Delta g| \leq 1}{\leq} e^{\varepsilon/2} \cdot P_T(t') \quad (3)$

- $P_Q(\underbrace{v' - \Delta g - \Delta f}_{\leq 2}) \leq e^{\varepsilon/2} \cdot P_Q(v') = e^{\varepsilon/2} \cdot P_{4/\varepsilon}(v').$ (Proven similarly!)

Combine ineq. (3) with (1):

$$(1) \leq \iint e^{\varepsilon/2} \cdot P_T(t') \cdot e^{\varepsilon/2} \cdot P_Q(v') \cdot \mathbb{1}\{g_{x'} < t' < f_k(x') + v'\} dt' dv'.$$

$$= e^\varepsilon \cdot \Pr[A(x') = a \mid \{v_i\}_{i=1}^{k-1}].$$

Integrate over $v_1, \dots, v_k$, and $a \in S$ $\blacksquare$.

Definition of accuracy guarantees. (A)

For a fixed prefix of k queries, I want:

$$a_i = \text{YES} \Rightarrow f_i(x) \geq T - \alpha$$

$$a_i = \text{NO} \Rightarrow f_i(x) \leq T + \alpha$$

Additionally, if $f_i(x) < T - \alpha$ for every $i \leq k$, do not stop before reaching k .

Accuracy: AboveThreshold satisfies the accuracy guarantee with

$$\alpha = \frac{\epsilon}{\epsilon} \left(\log k + \log \frac{2}{\delta} \right)$$

mild dependence on #iterations

w.p. $1 - \delta$.

Proof sketch:

Let event $E = \left\{ | \eta | < \frac{\alpha}{2}, \max_{i \in [k]} |v_i| < \frac{\alpha}{2} \right\}$. Under E : $a_i = \text{YES} \Rightarrow f_i(x) + v_i \geq \hat{T}$

$$\Leftrightarrow f_i(x) + v_i \geq T + \eta \Leftrightarrow f_i(x) - T \geq \eta - v_i \geq -\frac{\alpha}{2} - \frac{\alpha}{2} = -\alpha \Leftrightarrow f_i(x) \geq T - \alpha.$$

Similarly for $a_i = \text{NO}$. Trivially, also AboveThreshold does not halt before k if $f_i(x) < T - \alpha$ $\forall i \in [k]$.

$$\Pr[E^c] \leq \Pr[|\eta| \geq \frac{\alpha}{2}] + k \cdot \Pr[|v_i| \geq \frac{\alpha}{2}] \leq e^{-\epsilon \alpha / 4} + k e^{-\epsilon \alpha / 8} \leq \delta \text{ by the choice of } \alpha.$$

Sparse Vector Technique

What if we want to find c above-threshold values?

$\text{Sparse}(x, \{f_i\}, T, c, \epsilon_0)$

1. Set $\text{count} = 0$
2. Draw $\hat{T}_0 = T + \text{Lap}(\frac{2}{\epsilon_0})$.
3. For each $i = 1, 2, \dots$

Draw $v_i \sim \text{Lap}(\frac{4}{\epsilon_0})$

If $f_i(x) + v_i > \hat{T}_{\text{count}}$:

output $a_i = \text{YES}$ and $\text{count}++$

If $\text{count} = c$, halt.

Otherwise draw $\hat{T}_{\text{count}} = T + \text{Lap}(\frac{2}{\epsilon_0}) \leftarrow \text{fresh noise}$

Otherwise output $a_i = \text{NO}$ and continue.

Privacy: Notice that Sparse is an adaptive composition of c independent calls (no shared randomness) to AboveThreshold. So it's

$$\begin{cases} \epsilon\text{-DP} & \text{if } \epsilon_0 = \frac{\epsilon}{c} \\ (\epsilon, \delta)\text{-DP} & \text{if } \epsilon_0 = \frac{\epsilon}{\sqrt{8c \log(1/\delta)}} \end{cases}$$

← Again privacy depends on c .

Accuracy

Assume $L_\alpha(\{i: |i| \leq k: f_i(x) \geq T - \alpha\}) \leq c$. Then all outputs a_i for $i: f_i(x) \geq T - \alpha$ satisfy the accuracy guarantee (A) with

$$\alpha = \begin{cases} \frac{8c(\log k + \log(2c/\delta))}{\epsilon}, & \text{if } \delta = 0 \\ \frac{\sqrt{512c \cdot \log(1/\delta)}(\log k + \log(2c/\delta))}{\epsilon}, & \text{if } \delta > 0 \end{cases}$$

w.p. $1 - \delta$.

Remarks:

- If we want to release the values of some of the queries privately, we need to do this with fresh noise $f_i(x) + z_i$, $z_i \sim \text{Lap}(1/\epsilon)$ and account for it by composition. AboveThreshold relies on $(f_i(x) + v_i)$'s not being released for its privacy guarantee! So we can't release these directly.
- If all $f_i = f$, AboveThreshold can be seen as a randomized post processing of $Y = f(x) - \eta = f(x) + \text{Lap}(2/\epsilon)$ which is $(\epsilon/2)$ -DP.
- In the context of selection, think of $f_i(x) = q(y_i, x)$ as a score function of candidate y_i . The Exp.Mech. and Report-Noisy-max pick best y_i among k options with accuracy $\geq \text{OPT} - O(\frac{\log k}{\epsilon})$. AboveThreshold does the same in the online setting where each next y_i is chosen adaptively (with the caveat that it assumes knowledge of $T \approx \text{OPT}$ in this case).
- Recent work with sparse flavor: Cohen-Lyu'23: Target charging Technique for privacy analysis across interactive computations.