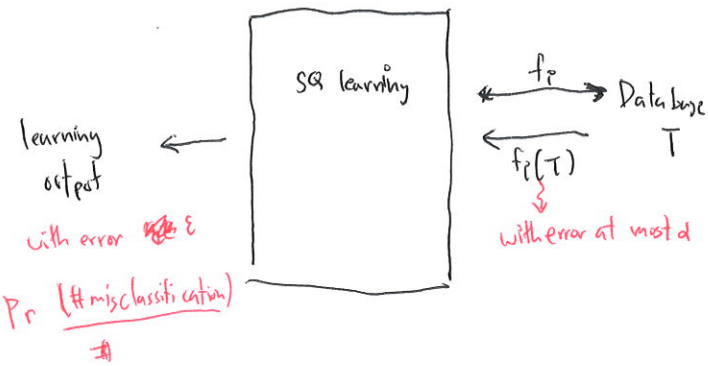
[Kwark and Roth, Chapter 11]

## Statistical Query Learning (SQ learning)

### Statistical Query (SQ)




$$f_i(T) = \sum_j \frac{f_i(T_j)}{\|T\|}$$

The output of small DB algorithm could be an input of SQ learning.

The error $\leq \alpha$ is obtained with prob. $1-\beta$.

$\Pr[\#\text{misclassification} / \#] \leq \varepsilon$

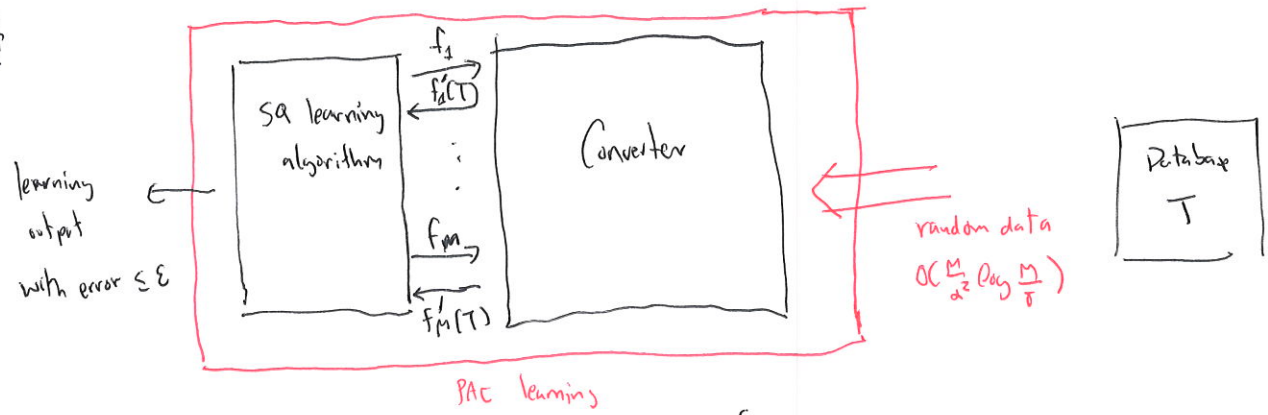
**Definition** An algorithm is SQ learning if, by $n = \text{poly}(1/\varepsilon)$ queries, we have a learning result with at most $\varepsilon$ errors.

**Theorem** We can obtain a PAC learning algorithm from an SQ learning algorithm.

In particular, if the SQ learning algorithm needs $M$ queries, we have a PAC learning with $O\left(\frac{M}{\alpha^2}\log\left(\frac{M}{\delta}\right)\right)$.

prob. that we have a learning output is prob. error $\leq \varepsilon$ in PAC learning

**Proof**




The converter will use $m = \frac{1}{\alpha^2}\log\frac{M}{\delta}$ for each queries $f_i$.

Suppose that we have $T_{j_1}, \ldots, T_{j_m}$ for query $i$.

$S\left(f_i'(T)\right) := \sum_k \frac{f_i(T_{j_k})}{m} X_k$ with expected value $f_i(T)$

$$\Pr\left[|f_i'(T) - f_i(T)| > \alpha\right] \leq 2\cdot\exp(-2\cdot\alpha^2 m^2)$$

($f_i'(T) = S$, $f_i(T) = \mu$)

$$= 2\cdot\exp\left(-2\cdot\alpha^2\cdot\frac{1}{\alpha^2}\log\frac{M}{\delta}\right)$$

$$= 2\cdot\left[\exp\left(\log\frac{M}{\delta}\right)\right]^{-2}$$

$$= 2\cdot\frac{M^{-2}}{\delta^{-2}} = 2\cdot\frac{\delta^2}{M^2} \leq \frac{\delta}{M}$$

when $\frac{\delta}{M} \leq 0.5$, which is true when $M \geq 2$

$$\Pr\left[|f_i'(T) - f_i(T)| \leq \alpha\right] \geq 1 - \frac{\delta}{M}$$

$$\Pr[\text{For some } i,\ |f_i'(T) - f_i(T)| > \alpha] \le \sum_i \underbrace{\Pr[|f_i'(T) - f_i(T)| > \alpha]}_{\delta/M} \le M \cdot \frac{\delta}{M} = \delta$$

$$\Pr[\text{For all } i,\ |f_i'(T) - f_i(T)| \le \alpha] \ge 1 - \delta.$$

Prob. that we have a valid input to the SQ learning algorithm

= Prob. that we have learning output with error $\le \varepsilon$. $= 1 - \delta$

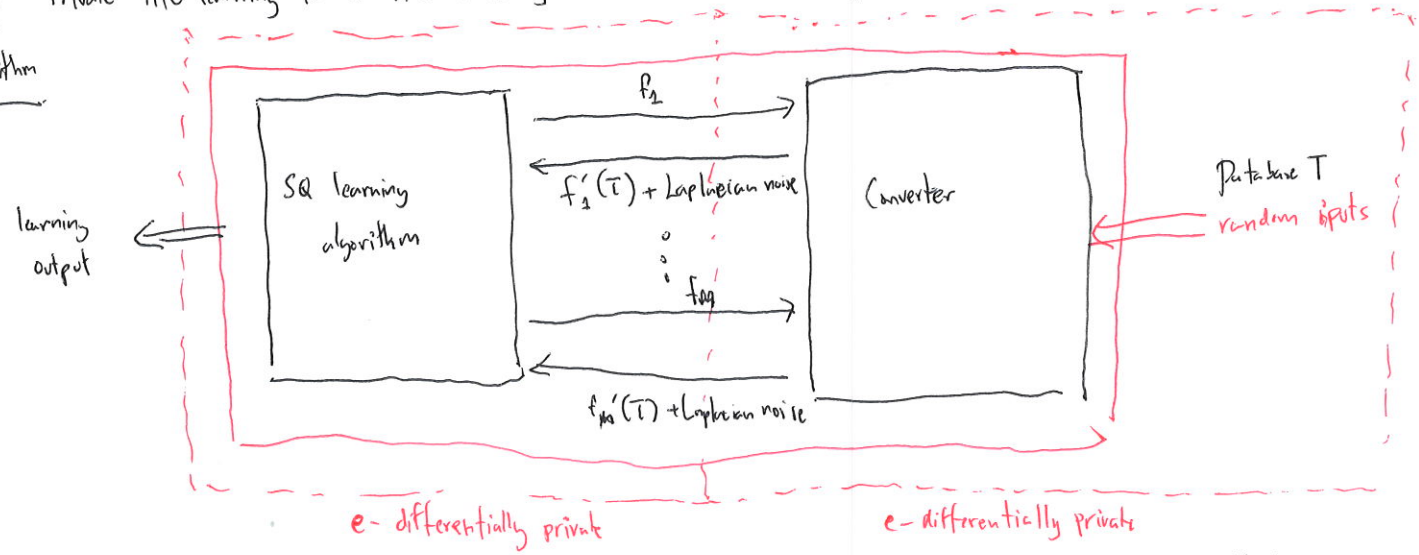
PAC learning

[1/6]

---

Recall Private PAC learning is a PAC learning that is $\varepsilon$-differentially private.

Algorithm




M publications $\varepsilon$-differentially private $\Rightarrow$ $\frac{\varepsilon}{M}$-differentially private for each publication

$\Rightarrow \text{Lap}\left(\frac{GS(f)}{\varepsilon}\right)$ — $GS(f)$: 1 because of the function $f$; $\varepsilon$: $\varepsilon/M$ → # queries use for each $f_i$

$\Rightarrow \text{Lap}\left(\frac{1}{\varepsilon} \cdot \frac{M}{\varepsilon}\right)$

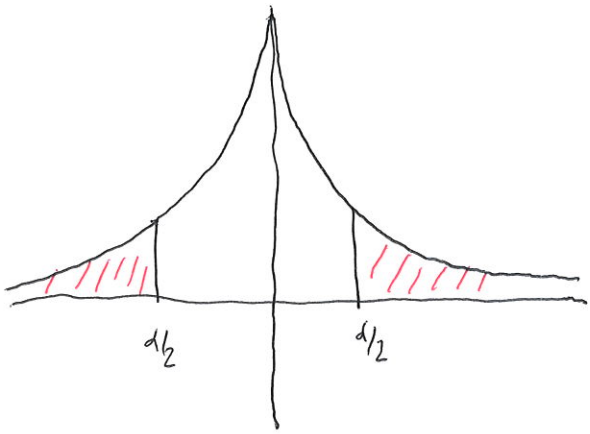
$b$

We will have $|f_i(T) - f_i'(T)| \le \alpha/2$ and

$|\text{Laplacian noise}| \le \alpha/2$

with probability $1 - \delta$,

Laplacian distribution




$$\Pr[\text{Noise} > \alpha/2] = 2 \cdot \exp\left(-\frac{\alpha}{2} \cdot \frac{\varepsilon}{1}\right)$$

$$= 2\exp\left(-\alpha/2 \cdot \frac{\varepsilon}{1} / M\right)$$

**Theorem** We have private PAC learning when $p = 2\left\{\frac{M}{\varepsilon\alpha} + \frac{1}{\alpha^2}\right\} \cdot \log\frac{4M}{\delta}$.

↑ additional term.

**Proof** $\Pr[|f_i'(T) - f_i(T)| > \alpha/2] = 2 \cdot \exp\left(-2 \cdot p \cdot \frac{\alpha^2}{4}\right)$

$$= 2 \cdot \exp\left(-2 \cdot 2\left\{\frac{M}{\varepsilon\alpha} + \frac{1}{\alpha^2}\right\} \cdot \log\frac{4M}{\delta} \cdot \frac{\alpha^2}{4}\right)$$

$p$

$$\leq 2 \cdot \exp\left(-\frac{1}{\alpha^2}\log\frac{4M}{\delta} \cdot \alpha^2\right)$$

$$= 2 \cdot \left[\exp\left(\log\frac{4M}{\delta}\right)\right]^{-1}$$

$$= 2 \cdot \frac{\delta}{4M} = \frac{\delta}{2M}$$

$$\Pr[|f_i'(T) - f_i(T)| > \alpha/2 \text{ for some } i] \leq \frac{\delta}{2M} \cdot M = \frac{\delta}{2}$$

$$\Pr[\text{Noise to } f_i > \alpha/2] = 2\exp\left(-\frac{\alpha\varepsilon}{2} \cdot \frac{1}{2\left\{\frac{1}{\varepsilon\alpha} + \frac{1}{\alpha^2}\right\}\log\frac{4M}{\delta}}\right)$$

$$= 2\exp\left(-\frac{\alpha\varepsilon}{2} \cdot 2\left\{\frac{M}{\varepsilon\alpha} + \frac{1}{\alpha^2}\right\} \cdot \log\frac{4M}{\delta} / M\right)$$

$$\leq 2 \cdot \exp\left(-\frac{\alpha\varepsilon}{2} \cdot \frac{2}{\varepsilon\alpha} M \log\frac{4M}{\delta} / M\right)$$

$$= 2 \cdot \frac{\delta}{4M} = \frac{\delta}{2M}$$

$$\Pr[\text{Noise to } f_i > \alpha/2 \text{ for some } i] \leq \frac{\delta}{2}$$

$$\Pr[|f_i'(T) - f_i(T)| > \alpha/2 \text{ or Noise to } f_i > \alpha/2 \text{ for some } i] \leq \delta/2 + \delta/2 = \delta$$

$$\Pr[|f_i'(T) - f_i(T)| \leq \alpha/2 \text{ and Noise to } f_i \leq \alpha/2 \text{ for all } i] \geq 1 - \delta.$$

$$\Pr[|(f_i'(T) + \text{Noise}) - f_i(T)| \leq \alpha \text{ for all } i] \geq 1 - \delta$$

Schedule from next week, ...

November 6 th  Midterm Exam. (30%)  Open book !!!

[Laplacian and Exponential mechanism, Small DB algorithm]

November 13 th  No Class. (for canceled classes at other days)

November 20 th  Optional Class (Introduction to Abstract Algebra)

November 27 th  Class 6 (Calculation on Elliptic Curve)

---

## Linking Attack

Data published by governments

| Name | Age | Hometown |
|---|---|---|
| Alice | 20 | Tokyo |
| Bob | 25 | Tokyo |
| Charles | 30 | Kyoto |
| Doc | 30 | Osaka |
| Eve | 35 | Osaka |

Data at Hospitals → Data Scientists

| Age | Hometown | Diabetes |
|---|---|---|
| 20 | Tokyo | ✓ |
| 25 | Tokyo | × |
| 30 | Kyoto | ✓ |
| 30 | Osaka | ✓ |
| 35 | Osaka | × |

- Alice has diabetes

---

## k-anonymity

1. put data into groups of k or larger.
2. make data in the same group looks indistinguishable by removing some information

| Hometown | Age | Diabetes |
|---|---|---|
| Tokyo | ~~20~~ | ✓ |
| Tokyo | ~~25~~ | × |
| ~~Kyoto~~ | ~~30~~ | ✓ |
| ~~Osaka~~ | ~~30~~ | ✓ |
| ~~Osaka~~ | ~~35~~ | × |

→ First group

→ Second group.

→

| Hometown | Age | Diabetes |
|---|---|---|
| Tokyo | * | ✓ |
| Tokyo | * | × |
| * | * | ✓ |
| * | * | ✓ |
| * | * | × |

we know that Alice is in this group, but we do not know which rows Alice is.

generalization

| Hometown | Age | Diabetes |
|---|---|---|
| Tokyo | 20-25 | ✓ |
| Tokyo | 20-25 | × |
| Kansai | 30-35 | ✓ |
| Kansai | 30-35 | ✓ |
| Kansai | 30-35 | × |

Alice

We can publish more information by generalization!

k is usually around 20 - 30 records

- large k → more privacy, reveal less information

## How to find the best grouping?

- No current solution. Clustering to have the similar data in all the groups.
- $\left(\frac{\|T\|}{k}\right)$-means ? — $O(k)$-approximation algorithm.

---

## $\ell$-diversity

| Hometown | Age | Diabetes |
|---|---|---|
| Tokyo | 20-25 | ✓ |
| Tokyo | 20-25 | ✓ |
| Kansai | 30-35 | ✗ |
| Kansai | 30-35 | ✗ |
| Kansai | 30-35 | ✓ |

← Alice is any of this two
But both of the two have diabetes.
Alice has diabetes! (:(

Suppose that records in a particular group $G_i$ are $\{T_{i_1}, \ldots, T_{i_k}\}$ with sensitive information

$S_i = \{s_{i_1}, \ldots, s_{i_k}\}$. # same entries in $S_i$ must be $\leq \frac{|G_i|}{\ell}$.

sensitive information: information which is important for data scientists, cannot be hidden, but users do not want to reveal.

In all groups, # of persons with diabetes should not be larger than

$$\frac{\text{group size}}{\ell}.$$

$\ell$ is usually around 5-6 records. large $\ell$ → more privacy, reveal less information

---

## Problem with $\ell$-diversity

| | Age | Salary |
|---|---|---|
| $G_1$ | 29 | 3M |
| | 22 | 4M |
| | 27 | 5M |
| $G_2$ | 43 | 6M |
| | 52 | 11M |
| | 47 | 8M |
| $G_3$ | 30 | 7M |
| | 36 | 9M |
| | 32 | 10M |

No sensitive information is not same in all groups

⇒

| Age | Salary |
|---|---|
| 22-29 | 3M |
| 22-29 | 4M |
| 22-29 | 5M |
| 43-52 | 6M |
| 43-52 | 11M |
| 43-52 | 8M |
| 30-36 | 7M |
| 30-36 | 9M |
| 30-36 | 10M |

Alice is in this group
We know that Alice's salary is small.

Before publication

Alice Salary $\in \{3M, 4M, \dots, 11M\}$ $\Longrightarrow$ After publication

Alice Salary $\in \{3M, 4M, 5M\}$

gain a lot of information

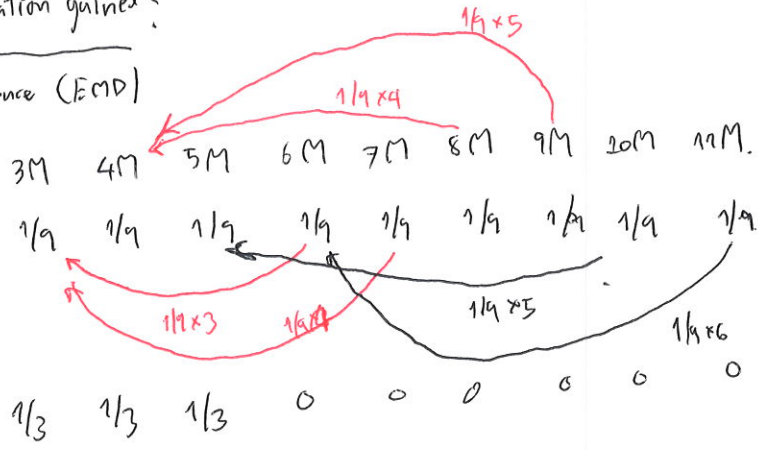
does not gain a lot of information

After publication Alice Salary $\in \{3M, 8M, 11M\}$

---

How to define information gained?

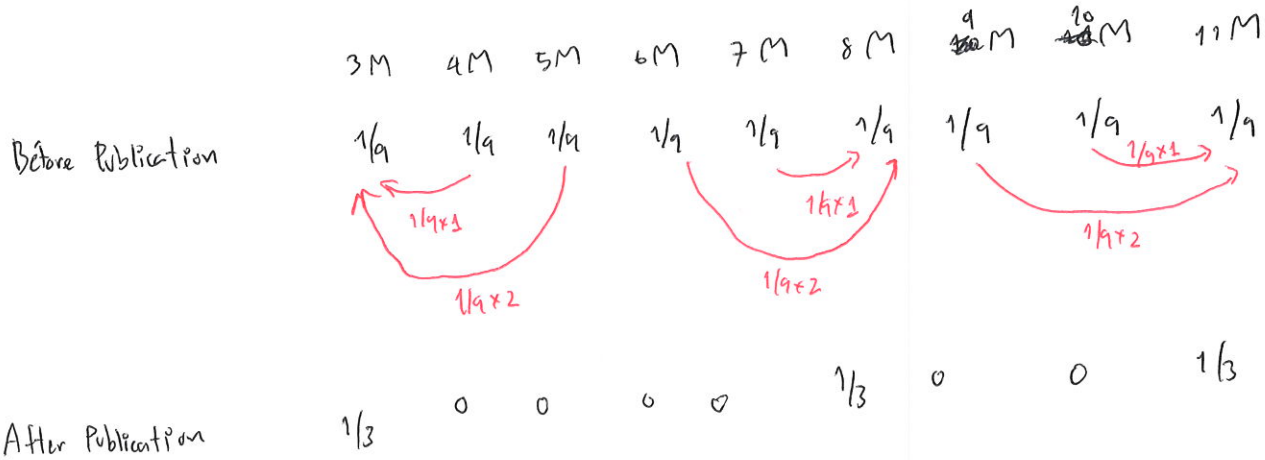
Earth Moving Distance (EMD)




Work done = $1/9 \times 3 + 1/9 \times 4 + 1/9 \times 4 + 1/9 \times 5 + 1/9 \times 5 + 1/9 \times 6 = 3$

EMD = ~~min~~ work done at the move with minimum work

EMD = 3.




EMD = $1/9 \times 1 + 1/9 \times 2 + 1/9 \times 2 + 1/9 \times 1 + 1/9 \times 2 + 1/9 \times 1 = 1$

More distance = more information leaked.

Information leaked by $\{3M, 4M, 5M\}$ is 3 times larger than that leaked by $\{3M, 8M, 11M\}$.

# t- closeness

Each group must have $EMD \leq t$

When $t=0$,

- All group contains exactly same sensitive information as a whole table.
  We cannot make sense of the published information.

When $t \to \infty$

- We cannot guarantee the privacy at all.

No standard $t$.