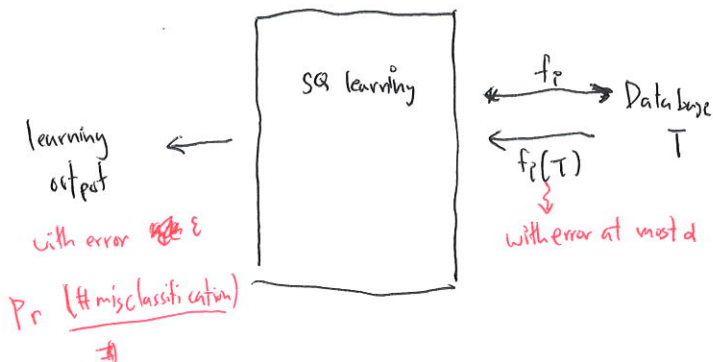


Statistical Query Learning (SQ learning)

Statistical Query (SQ)



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

The output of small DB algorithms could be an input of SQ learning,

The error $\leq d$ is obtained with prob. $1 - \delta$.

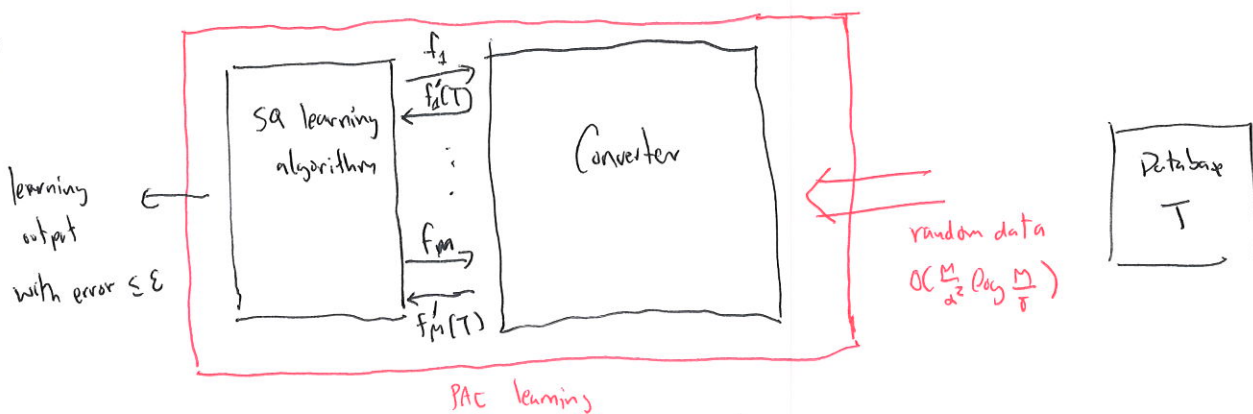
Definition An algorithm is SQ learning if, by $n = \text{poly}(1/\epsilon)$ queries, we have a learning result with at most ϵ errors.

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

In particular, if the SQ learning algorithm needs M queries, we ~~need~~ 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 $\leq \epsilon$ in PAC learning

Proof



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

Suppose that we have $T_{i,1}, \dots, T_{i,m_i}$ for query i .

$$S \left(\frac{f_i'(T)}{m_i} \right) = \sum_k \frac{f_i(T_{i,k})}{m_i} \quad \text{with expected value } f_i(T)$$

$$\begin{aligned} \Pr\left[\left| \frac{f_i'(T)}{m_i} - f_i(T) \right| > \alpha \right] &\leq 2 \cdot \exp(-2 \cdot \alpha^2 \cdot m_i) \\ &= 2 \cdot \exp(-2 \cdot \alpha^2 \cdot \frac{1}{\alpha^2} \log \frac{M}{\delta}) \\ &= 2 \cdot \left[\exp\left(\log \frac{M}{\delta}\right) \right]^{-2} \\ &= 2 \cdot \frac{M^{-2}}{\delta^{-2}} = 2 \cdot \frac{\delta^2}{M^2} \leq \frac{\delta}{M} \end{aligned}$$

when $\frac{\delta}{M} \leq 0.5$.

which is true when $M \geq 2/\delta$

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

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

$$\leq M \cdot \frac{\delta}{M} = \delta$$

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

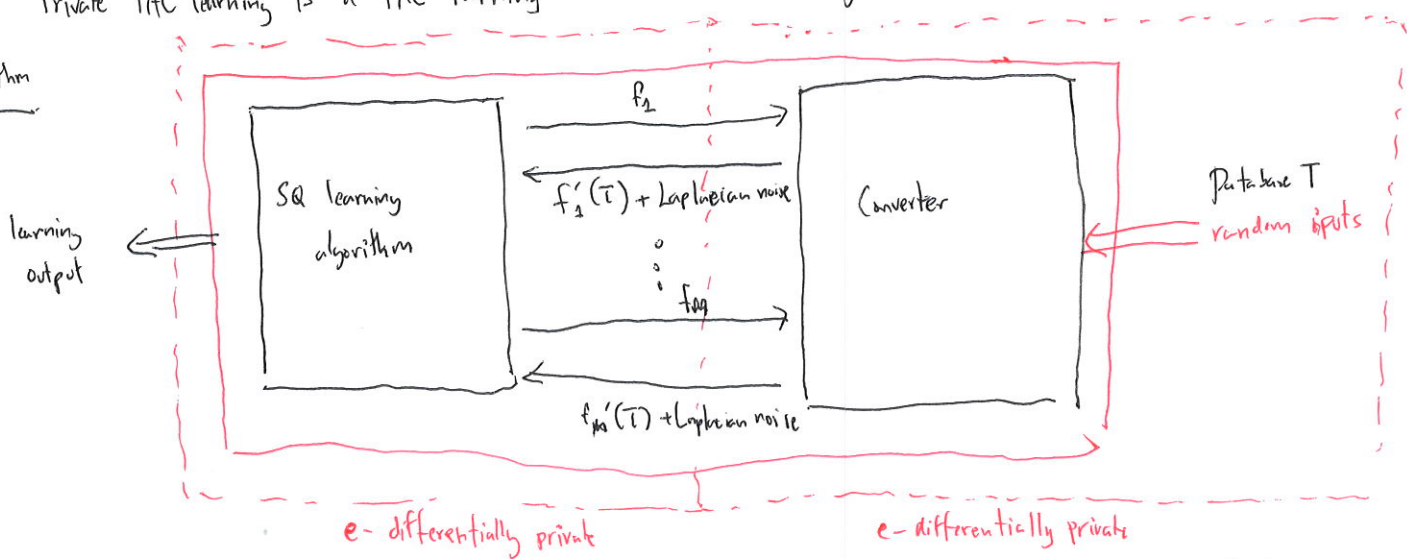
Prob. that we have a valid input to the SQ learning algorithm.

$$= \text{Prob. that we have learning output with error } \leq \epsilon. = 1 - \delta$$

PAC learning

Recall Private PAC learning is a PAC learning that is ϵ -differentially private.

Algorithm



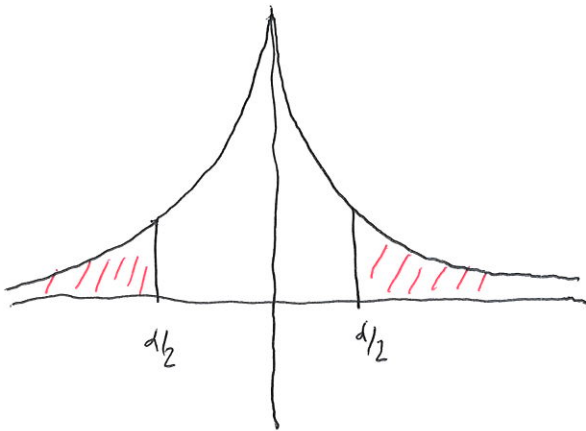
M publications ϵ -differentially private $\Rightarrow \frac{\epsilon}{M}$ ϵ -differentially private for each publication

$\Rightarrow \text{Lap}\left(\frac{b \cdot \epsilon(T)}{\epsilon}\right)$ because of the function f

$\Rightarrow \text{Lap}\left(\frac{1}{\frac{\epsilon}{M}}\right)$ # queries use for each f_i

We will have $|f_i(T) - f_i'(T)| \leq \alpha/2$ and $| \text{Laplacian noise} | \leq \alpha/2$ with probability $1 - \delta$.

Laplacean distribution



$$\Pr[\text{Noise} > d/2] = 2 \cdot \exp(-d/2 \cdot \frac{1}{b})$$

$$= 2 \exp(-d/2 \cdot \frac{e \cdot p}{1/M})$$

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

↑
additional term.

Proof

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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
Doe	30	Osaka
Eve	35	Osaka

Data at Hospitals → Data Scientists

Age	Hometown	Diabetes
20	Tokyo	✓
25	Tokyo	x
30	Kyoto	✓
30	Osaka	✓
35	Osaka	x

o 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	x
Kyoto	30	✓
Osaka	30	✓
Osaka	35	x

→ First group

→ Second group.

Hometown	Age	Diabetes
Tokyo	*	✓
Tokyo	*	x
*	*	✓
*	*	✓
*	*	x

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

generalization

Hometown	Age	Diabetes
Tokyo	20-25	✓
Tokyo	20-25	x
Kansai	30-35	✓
Kansai	30-35	✓
Kansai	30-35	x

) Alice

We can publish more information by generalization!

k is usually around

20-30 records

- large k → more privacy,

reveal less information