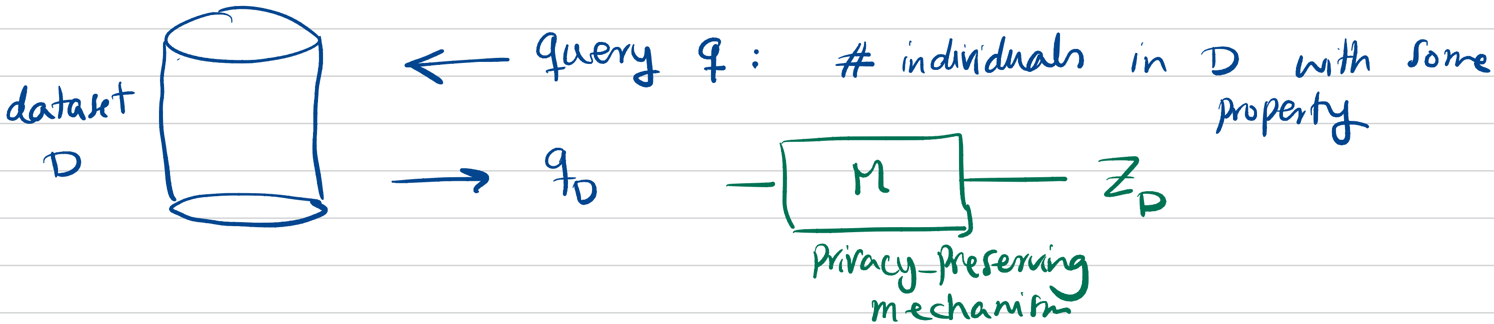
Chapter 2: Privacy

Privacy desideratum: The output of the algorithm shouldn't change an adversary's knowledge about a dataset **at all**.

This statement ensures that **no** information should be revelead about the dataset; which is very unrealistic & goes against the point of **learning**. A more realistic goal is:

Practical privacy desideratum: The output of the algorithm shouldn't significantly change an adversary's knowledge.

How to mathematically formulate this goal:



* **Goal:** We wish to design M in such a way that an adversary observing Z_D shouldn't be able to distinguish D from another dataset D' that differ in one entry!

dataset D

Name	Sex	Blood		HIV?
Chen	F	B		Y
Jones	M	A		N
Smith	M	O		N
Ross	M	O		Y
Lu	F	A		N
Shah	M	B		Y

$\rightarrow$ C (trusted "curator") $\leftarrow$ query q

C $\rightarrow$ q_D — M — $Z_D \sim M_D$

M: Privacy-Preserving mechanism

dataset D'

Name	Sex	Blood		HIV?
Chen	F	B		Y
Jones	M	A		N
Smith	M	O		N
~~Ross~~ Bob	M	O		~~Y~~ N
Lu	F	A		N
Shah	M	B		Y

$\rightarrow$ C (trusted "curator") $\leftarrow$ query q

C $\rightarrow$ $q_{D'}$ — M — $Z_{D'} \sim M_{D'}$

If M_D & $M_{D'}$ are "close" $\Rightarrow$ adversary can't distinguish D from D'

" $D \sim D'$ "

$\sim$: neighboring dataset

Jones has a plausible deniability

Def: Randomized mechanism M is said to be ε_DP for $\varepsilon \geq 0$

if

$$\bar{E}_{\gamma}(M_D \| M_{D'}) = 0 \quad \forall\, D \sim D'$$

$\gamma = e^{\varepsilon}$

neighboring datasets

* Since M_D & $M_{D'}$ are close, no adversary can distinguish M_D from $M_{D'}$ (& hence D from D').

Remarks:

1. Why HS divergence? It's known that TV doesn't lead to any meaningful privacy definition

(See lecture note of Salil Vadhan & also the 2nd assignment)

sec 1.6 "The complexity of differential privacy"

The main reason for HS divergence is its connection with hypothesis testing, making the DP definition operational.

2- DP was originally proposed by Dwork, McSherry, Nissim, & Smith in the paper "Calibrating noise to sensitivity in Private Data Analysis"

3- More Standard def:

M is ε-DP $\iff$ $M_D(A) \leq e^{\varepsilon} M_{D'}(A)$ $\forall$ interval A & $D \sim D'$

* The equivalence between this formula & the above

definition comes from the variational expression of HS divergence $E_\gamma(p\|q) = \sup_A \left[p(A) - \gamma q(A)\right]$

3_ You may even see some definition like this:

M is ε-DP $\iff$ $\underbrace{e^{-\varepsilon} M_{D'}(A) \leq}_{}$ $M_D(A) \leq e^{\varepsilon} M_{D'}(A)$ $\forall$ interval A & $D \sim D'$

This part is redundant! ← Prove this!

* To prove a mechanism M is ε-DP, it suffices to show

$E_{e^\varepsilon}(M_D \| M_{D'}) = 0$ or $\dfrac{M_D(A)}{M_{D'}(A)} \leq e^{\varepsilon}$ for any event A & any $D \sim D'$

4- the definition doesn't depend on the dataset that you have hand.

* Note that the above definition should hold for any arbitrary pair of D & D'

* **Perfect Privacy:** $\varepsilon = 0$

Let M be an ε-DP with $\varepsilon = 0$. Then since we have

M is ε-DP $\Leftrightarrow$ $E_{\underbrace{e^{\varepsilon}}_{=1}}(M_D \| M_{D'}) = 0$ $\quad \forall D \sim D'$

so $TV(M_D, M_{D'}) = 0 \Leftrightarrow M_D = M_{D'}$

So the distributions of Z_D & $Z_{D'}$ are exactly the same.

Great in terms of privacy, but leads to very poor utility.

$\varepsilon > 0$, but sufficiently small:

$$M \text{ is } \varepsilon\text{-DP} \implies \mathcal{E}_{e^{\varepsilon}}(M_D \| M_{D'}) = 0$$

$$\iff M_D(A) - e^{\varepsilon} M_{D'}(A) \leq 0 \quad \forall \text{ event } A$$

Since ε is small then $e^{\varepsilon} \approx 1 + \varepsilon$ so

$$M_D(A) - (1+\varepsilon) \cdot M_{D'}(A) \leq 0$$

$$\Rightarrow \quad M_D(A) - M_{D'}(A) \leq \varepsilon \; M_{D'}(A) \leq \varepsilon$$

$$\Rightarrow \quad \left| M_D(A) - M_{D'}(A) \right| \leq \varepsilon \quad \text{for any event } A.$$

Thus, for small ε, DP means no matter what value output takes, D & D' can't be distinguished.

$\varepsilon = \infty$.

$$\frac{M_D(A)}{M_{D'}(A)} \leq e^{\varepsilon}$$

; this requirement for large ε is always satisfied

ε characterizes the privacy gaurantee of mechanism M

The smaller ε is the better privacy is

Operational Privacy guarantee

We have a mechanism M that is ε-DP:

* Suppose M generates an output Z.

Null hypo: H_0: Dataset $= D$ ← Alice is in D

Alterative hyp: H_1: Dataset $= D'$ ← but not in D'

The goal is for an adversary to reliably test H_0 against H_1.

* Assumption: Adversary knows everyone else in real dataset.

* What does it mean to test H_0 against H_1 reliably?

Let's $\varphi : \mathcal{Z} \rightarrow \{0,1\}$

test function

$\varphi(Z) = 0 \longrightarrow H_0$ is accepted

$\varphi(Z) = 1 \longrightarrow H_1$ is accepted.

For any test function φ, we associate two errors:

1- Adversary rejects H_0 when H_0 was correct

"False positive" "Type I error"

FP

2- Adversary accepts H_0 when H_1 was correct

"False negative" "Type II error"

FN

When there exist a test function φ such that both FP & FN are small, we say that H_0 can be tested against H_1 reliably.

FP, FN ≈ 0

* claim: If Z is the output of an ε-DP mechanism with small ε, then H_0 can't be tested against H_1.

proof: Given φ, we can characterize FP & FN using the following set:

$$A = \{ z \in Z : \varphi(z) = 0 \}$$

$z \in A \Rightarrow H_0$ is accepted

$z \in A^c \Rightarrow H_1$ is accepted

Think about it!

$$FP = M_D(A^c) = 1 - M_D(A)$$

$$FN = M_{D'}(A)$$

$$FP + e^{\varepsilon} FN = 1 - M_D(A) + e^{\varepsilon} M_{D'}(A)$$

$$= 1 - \left[M_D(A) - e^{\varepsilon} M_{D'}(A) \right]$$

To find the best test function φ, we need to find φ resulting in smallest FP & FN, or equivalently the best set A:

$$\inf_A \; FP + e^{\varepsilon} FN = \inf_A \left[1 - \left[M_D(A) - e^{\varepsilon} M_{D'}(A) \right] \right]$$

$$= 1 - \sup_A \left[M_D(A) - e^{\varepsilon} M_{D'}(A) \right]$$

$$= 1 - E_{e^{\varepsilon}}(M_D \| M_{D'})$$

Since M is ε-DP, we have $E_{e^\varepsilon}(M_D \| M_{D'}) = 0$ & hence

$$\boxed{\inf \ FP + e^{\varepsilon} FN = 1}$$

theorem: If Z is the output of an ε-DP mechanism, then for **any** test function:

$$FP + e^{\varepsilon} FN \geq 1$$

$$FN + e^{\varepsilon} FP \geq 1$$

This theorem demonstrates that no test function can be found with FP & $FN \approx 0$ $\Rightarrow$ H_0 can't be tested against H_1 reliably!