Statistical Machine Learning

Privacy – Parts 01

(Version 1.0)

Hamid R. Rabiee

Spring 2023

Privacy Basics: What is Privacy?

- Privacy is the protection of an individual's personal information.

- Privacy is the rights and obligations of individuals and organizations with respect to the collection, use, retention, disclosure and disposal of personal information.

- Privacy $\neq$ Confidentiality

OECD Privacy Principles

1. Collection Limitation Principle

There should be limits to the collection of personal data and any such data should be obtained by lawful and fair means and, where appropriate, with the knowledge or consent of the data subject.

2. Data Quality Principle

Personal data should be relevant to the purposes for which they are to be used, and, to the extent necessary for those purposes, should be accurate, complete and kept up-to-date.

OECD Privacy Principles

3. Purpose Specification Principle

The purposes for which personal data are collected should be specified not later than at the time of data collection and the subsequent use limited to the fulfilment of those purposes or such others as are not incompatible with those purposes and as are specified on each occasion of change of purpose.

4. Use Limitation Principle

Personal data should not be disclosed, made available or otherwise used for purposes other than those specified in accordance with Principle 3 except:

a) with the consent of the data subject; or

b) by the authority of law.

OECD Privacy Principles

5. Security Safeguards Principle

Personal data should be protected by reasonable security safeguards against such risks as loss or unauthorized access, destruction, use, modification or disclosure of data.

6. Openness Principle

There should be a general policy of openness about developments, practices and policies with respect to personal data. Means should be readily available of establishing the existence and nature of personal data, and the main purposes of their use, as well as the identity and usual residence of the data controller.

OECD Privacy Principles

7. Individual Participation Principle

An individual should have the right:

a) to request to know whether or not the data controller has data relating to him;

b) to request data relating to him, …

c) to be given reasons if a request is denied; and

d) to request the data to be rectified, completed or amended.

8. Accountability Principle

A data controller should be accountable for complying with measures which give effect to the principles stated above.

Areas of Privacy

- Anonymity

  Anonymous communication:

  e.g., The TOR software to defend against traffic analysis
- Web privacy

  Understand/control what web sites collect, maintain regarding personal data
- Mobile data privacy, e.g., location privacy
- Privacy-preserving data usage

Privacy Preserving Data Sharing

The need to sharing data:

- For research purposes
  - E.g., social, medical, technological, etc.
- Mandated by laws and regulations:
  - E.g., census
- For security/business decision making:
  - E.g., network flow data for Internet-scale alert correlation
- For system testing before deployment:
- …

However, publishing data may result in privacy violations!

Privacy Basics: What is Privacy?

Example:

What information can be published?

Average height of US people

Height of an individual

Intuition:

If something is insensitive to the change of any individual tuple, then it should not be considered private

Example:

- Assume that we arbitrarily change the height of an individual in Iran
- The average height of Iranian people would remain roughly the same
- i.e., the average height reveals little information about the exact height of any particular individual

ε-Differential Privacy

Motivation:

It is OK to publish information that is insensitive to the change of any particular tuple in the dataset

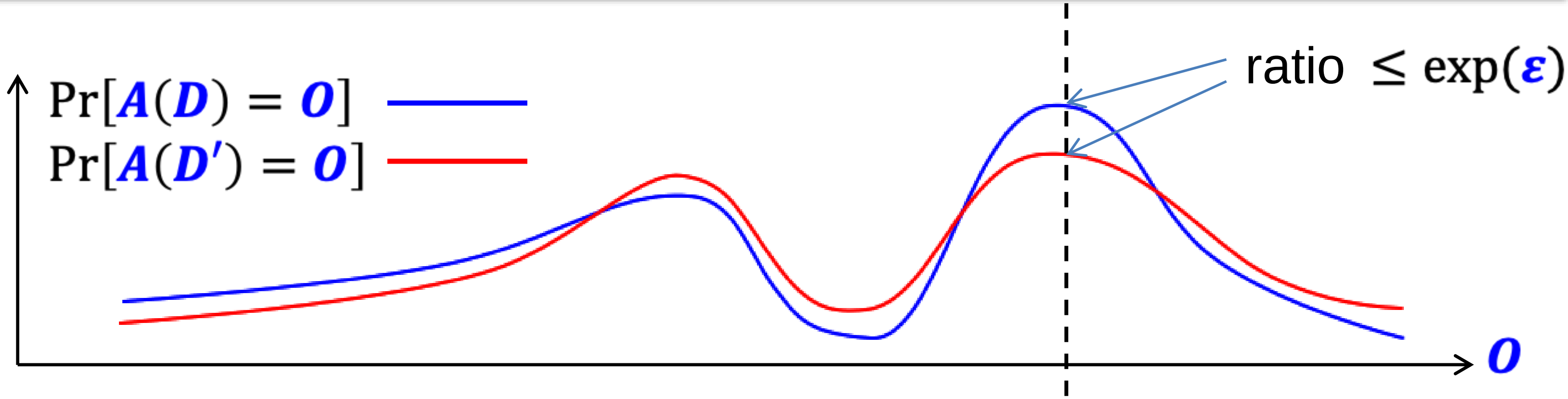
Definition:

Neighboring datasets: Two datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$, such that $\boldsymbol{D'}$ can be obtained by changing one single tuple in $\boldsymbol{D}$

A randomized algorithm $\boldsymbol{A}$ satisfies ε-differential privacy, iff for any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$ and for any output $\boldsymbol{O}$ of $\boldsymbol{A}$:

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \le \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

ε-Differential Privacy



Neighboring datasets: Two datasets D and D', such that D' can be obtained by changing one single tuple in D

A randomized algorithm A satisfies ε-differential privacy, iff for any two neighboring datasets D and D' and for any output O of A:

$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

The value of ε decides the degree of privacy protection

Achieving ε-Differential Privacy

Example:

Dataset: A set of patients

Objective: Release the number of diabetes patients with ε-differential privacy

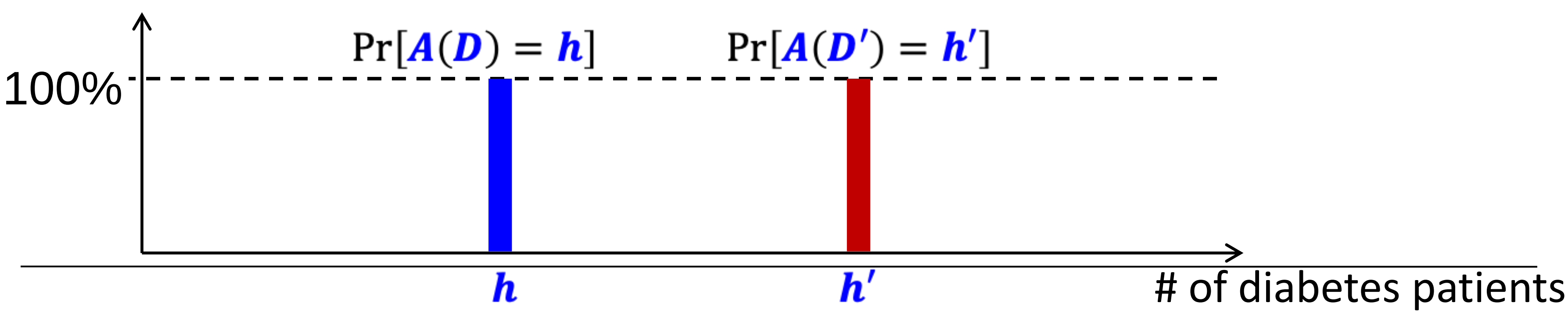
$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

It won't work if we release the number directly:

D : the original dataset

D': modify an arbitrary patient in D

$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$ does not hold for any ε



Achieving ε-Differential Privacy

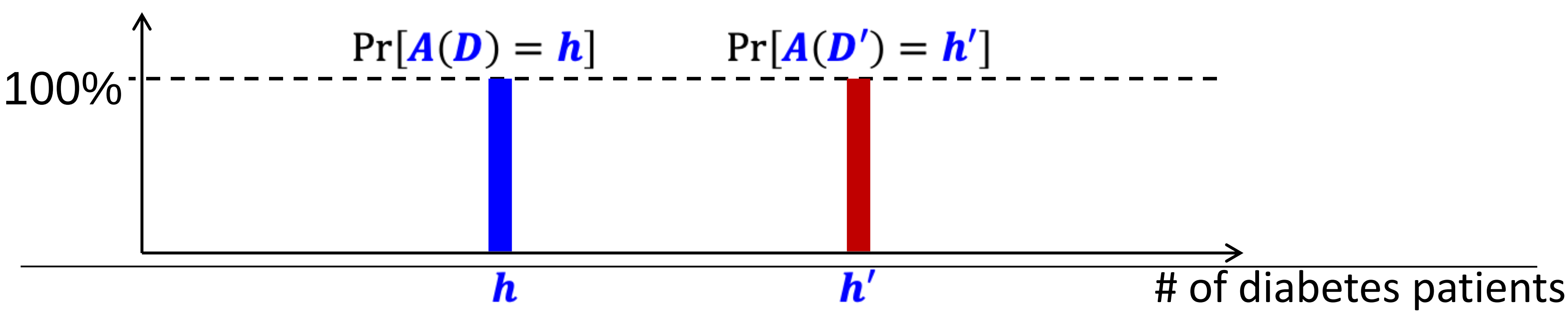
Example:

Dataset:	A set of patients
Objective:	Release the number of diabetes patients with ε-differential privacy

$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

Idea:

Perturb the number of diabetes patients to obtain a smooth distribution



Achieving ε-Differential Privacy

Example:

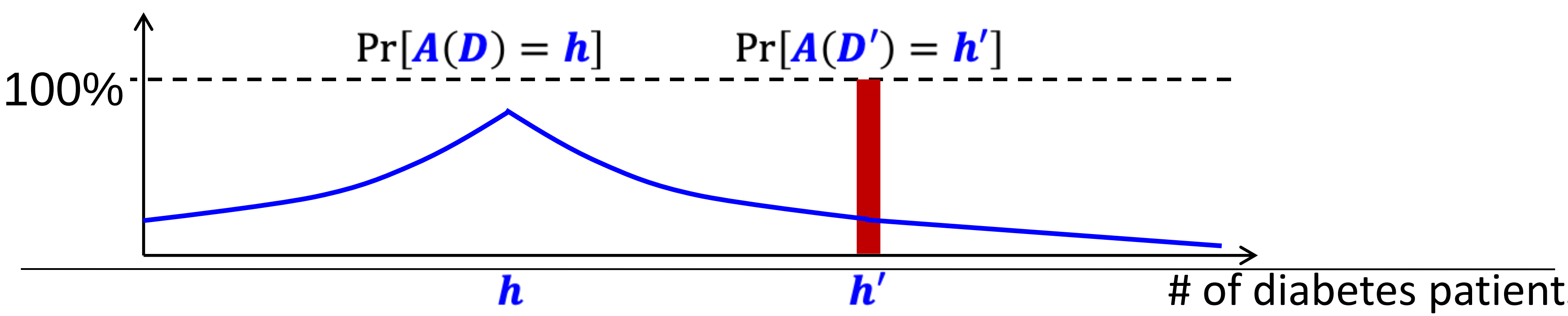
Dataset: A set of patients

Objective: Release the number of diabetes patients with ε-differential privacy

$$\Pr[A(D) = O] \leq \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

Idea:

Perturb the number of diabetes patients to obtain a smooth distribution



Achieving $\boldsymbol{\varepsilon}$-Differential Privacy

Example:

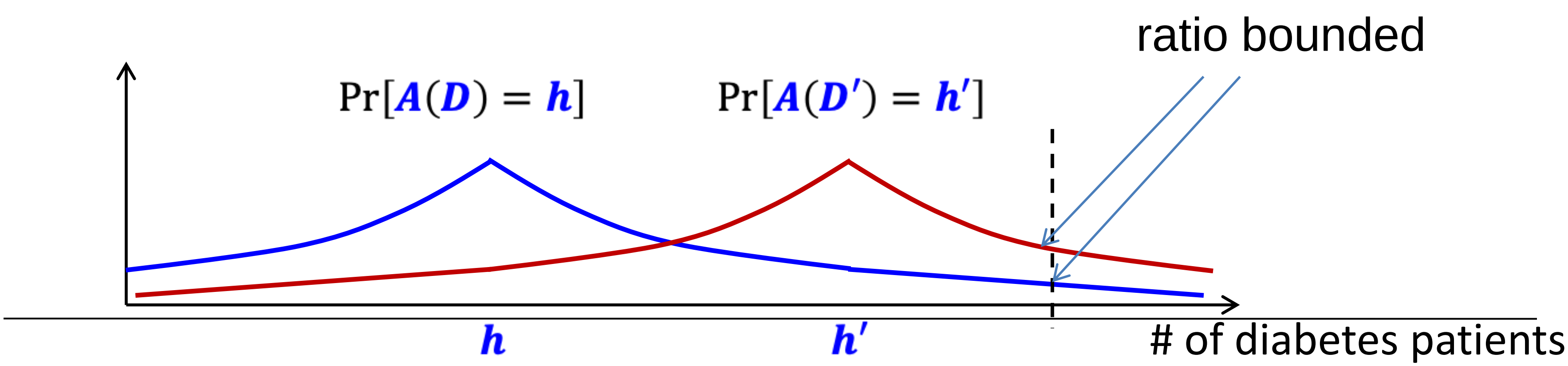
Dataset: A set of patients

Objective: Release the number of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \le \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

Idea:

Perturb the number of diabetes patients to obtain a smooth distribution



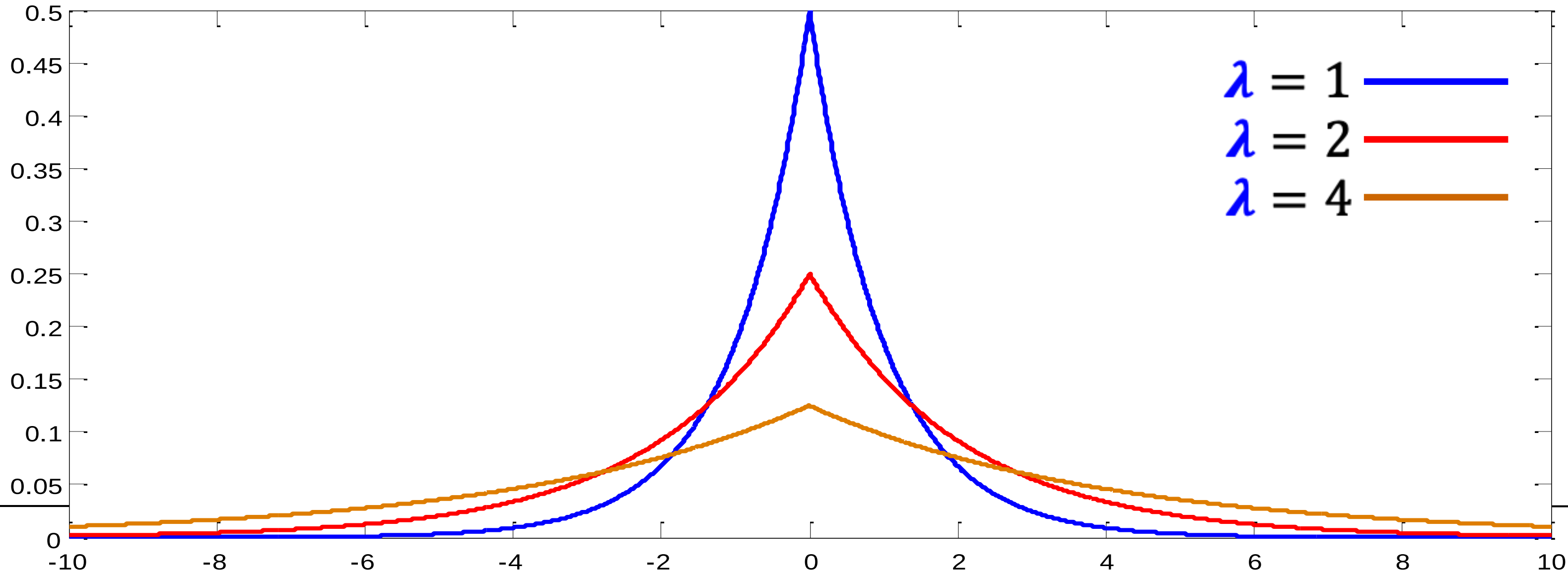
Laplace Distribution

$$pdf(\boldsymbol{x}) = \frac{1}{2\boldsymbol{\lambda}}\exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right);$$

increase/decrease $\boldsymbol{x}$ by $\boldsymbol{\alpha}$

→ $pdf(\boldsymbol{x})$ changes by a factor of $\exp\left(-\frac{|\boldsymbol{\alpha}|}{\boldsymbol{\lambda}}\right)$

variance: $2\boldsymbol{\lambda}^2$; $\boldsymbol{\lambda}$ is referred as the *scale*



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with ε-differential privacy

$$\Pr[A(D) = O] \le \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

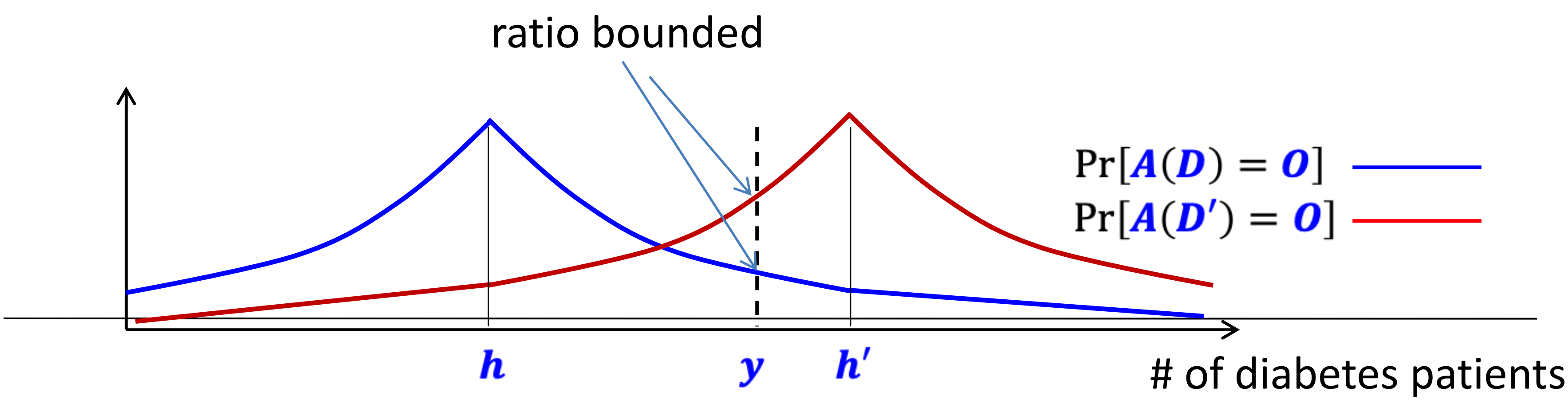
Method: Release the number + Laplace noise

$$pdf(x) = \exp\left(-\frac{|x|}{\lambda}\right)/2\lambda$$

Rationale:

D : the original dataset; # of diabetes patients = h

D': modify a patient in D; # of diabetes patients = h'



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

Method: Release the number + Laplace noise

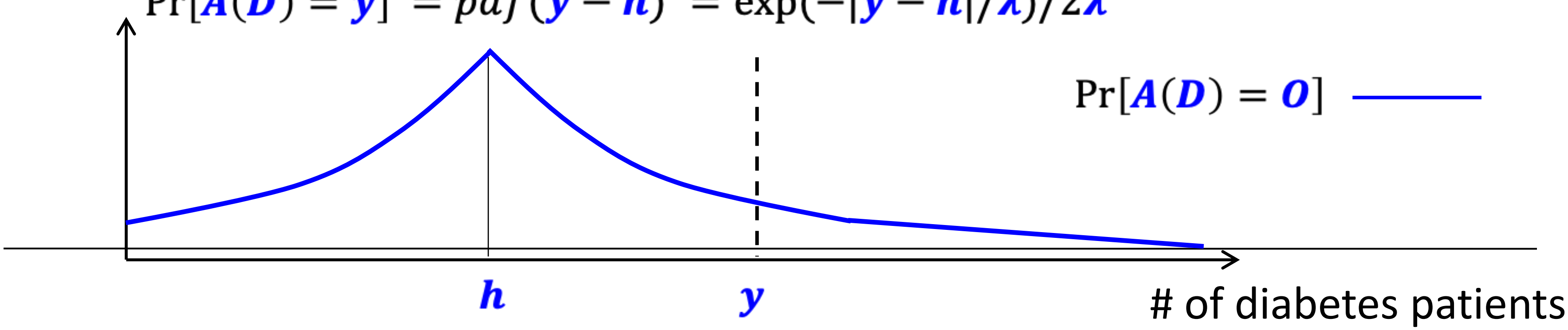
$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

Rationale:

$\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$

$\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}) = \exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

Method: Release the number + Laplace noise

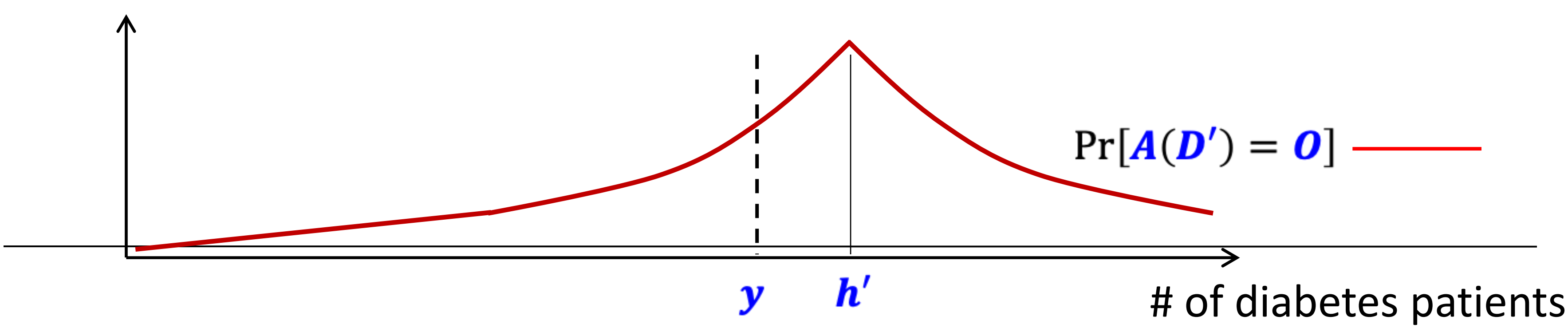
$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

Rationale:

$\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$

$\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}] = pdf(\boldsymbol{y} - \boldsymbol{h}') = \exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}$$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with ε-differential privacy

$$\Pr[A(D) = O] \le \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

Method: Release the number + Laplace noise

$$pdf(x) = \exp\left(-\frac{|x|}{\lambda}\right)/2\lambda$$

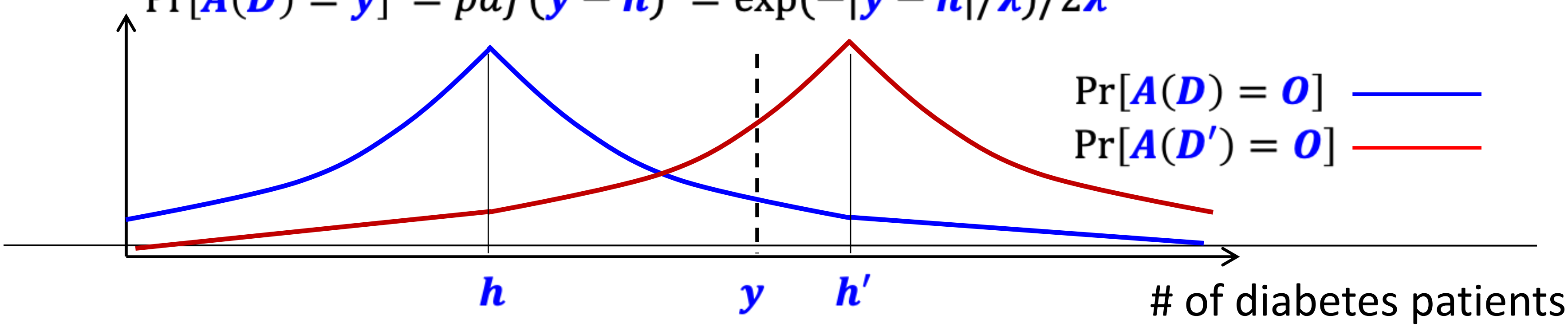
Rationale:

D : the original dataset; # of diabetes patients = h

D': modify a patient in D; # of diabetes patients = h'

$$\Pr[A(D') = y] = pdf(y - h') = \exp(-|y - h'|/\lambda)/2\lambda$$

$$\Pr[A(D) = y] = pdf(y - h) = \exp(-|y - h|/\lambda)/2\lambda$$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

Method: Release the number + Laplace noise

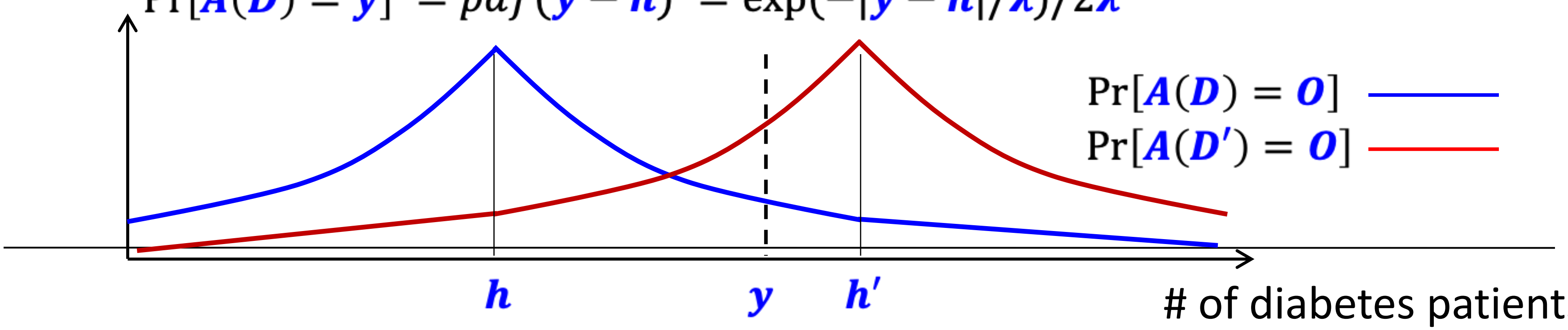
$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

Rationale:

$\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$

$\boldsymbol{D'}$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h'}$

$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} \quad \begin{matrix} = pdf(\boldsymbol{y} - \boldsymbol{h'}) = \exp(-|\boldsymbol{y} - \boldsymbol{h'}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda} \\ = pdf(\boldsymbol{y} - \boldsymbol{h}) = \exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda} \end{matrix}$$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with ε-differential privacy

$$\Pr[A(D) = O] \le \exp(\varepsilon) \cdot \Pr[A(D') = O]$$

Method: Release the number + Laplace noise

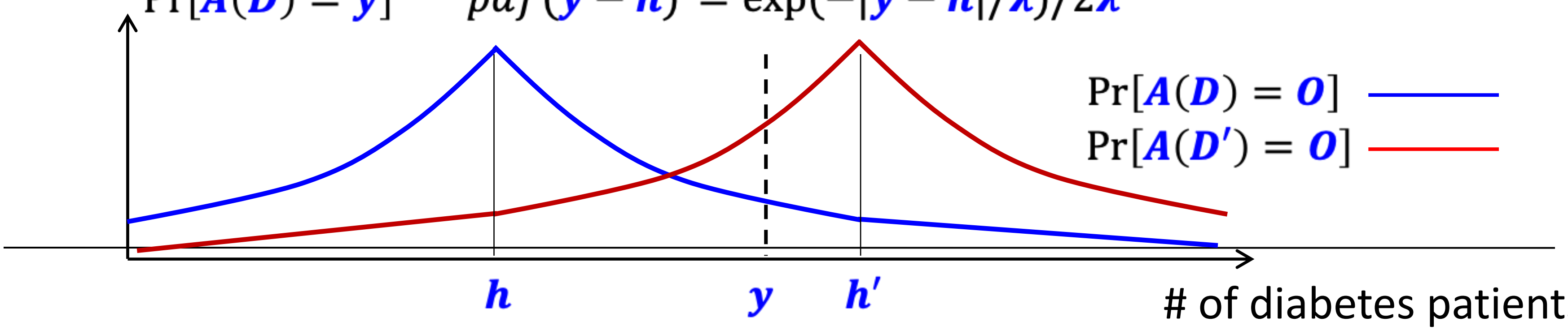
$$pdf(x) = \exp\left(-\frac{|x|}{\lambda}\right)/2\lambda$$

Rationale:

D : the original dataset; # of diabetes patients = h

D': modify a patient in D; # of diabetes patients = h'

$$\frac{\Pr[A(D') = y]}{\Pr[A(D) = y]} = \frac{pdf(y - h')}{pdf(y - h)} \begin{matrix} = \exp(-|y - h'|/\lambda)/2\lambda \\ = \exp(-|y - h|/\lambda)/2\lambda \end{matrix}$$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \le \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

Method: Release the number + Laplace noise

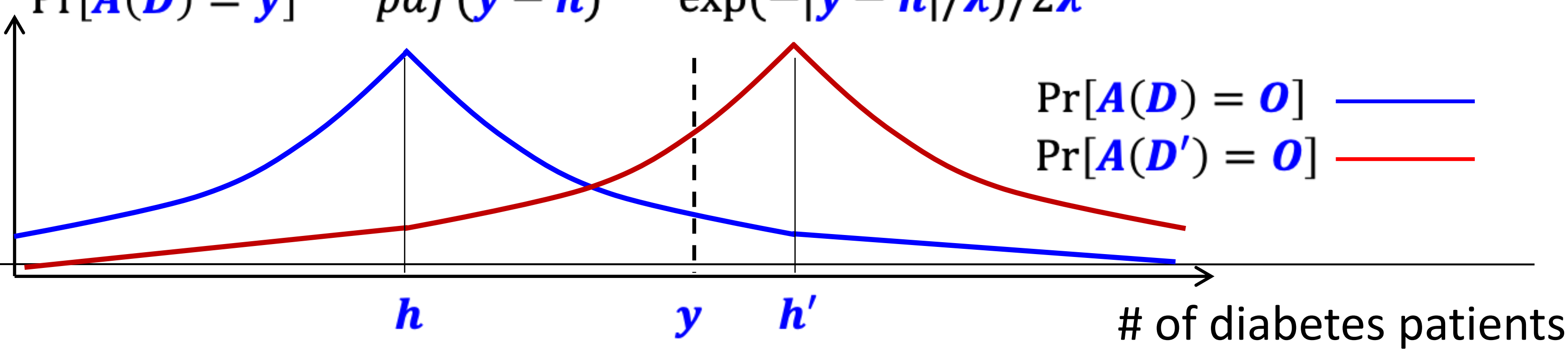
$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

Rationale:

$\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$

$\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} = \frac{pdf(\boldsymbol{y} - \boldsymbol{h}')}{pdf(\boldsymbol{y} - \boldsymbol{h})} = \frac{\exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}}{\exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})/2\boldsymbol{\lambda}}$$



Differential Privacy via Laplace Noise

Dataset: A set of patients

Objective: Release # of diabetes patients with $\boldsymbol{\varepsilon}$-differential privacy

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \le \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{O}]$$

Method: Release the number + Laplace noise

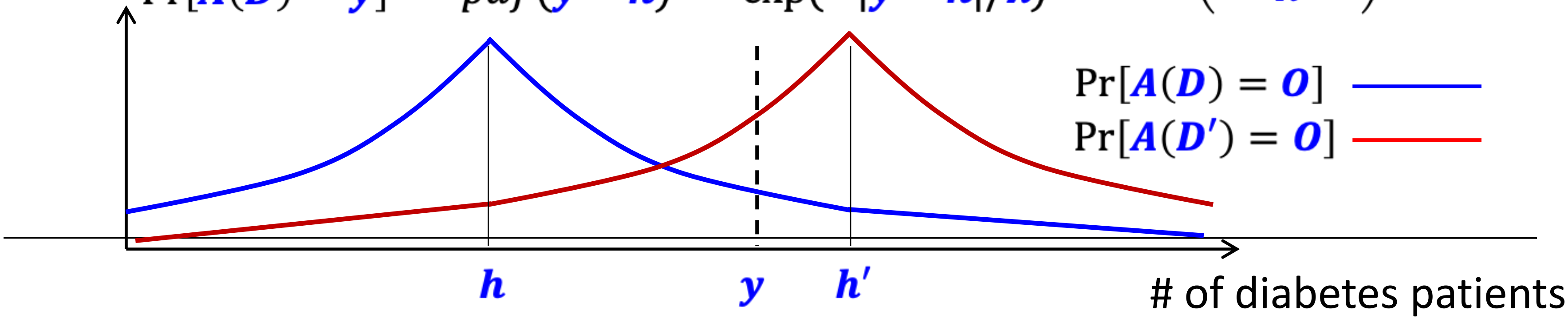
$$pdf(\boldsymbol{x}) = \exp\left(-\frac{|\boldsymbol{x}|}{\boldsymbol{\lambda}}\right)/2\boldsymbol{\lambda}$$

Rationale:

$\boldsymbol{D}$: the original dataset; # of diabetes patients = $\boldsymbol{h}$

$\boldsymbol{D}'$: modify a patient in $\boldsymbol{D}$; # of diabetes patients = $\boldsymbol{h}'$

$$\frac{\Pr[\boldsymbol{A}(\boldsymbol{D}') = \boldsymbol{y}]}{\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{y}]} = \frac{pdf(\boldsymbol{y} - \boldsymbol{h}')}{pdf(\boldsymbol{y} - \boldsymbol{h})} = \frac{\exp(-|\boldsymbol{y} - \boldsymbol{h}'|/\boldsymbol{\lambda})}{\exp(-|\boldsymbol{y} - \boldsymbol{h}|/\boldsymbol{\lambda})} \le \exp\left(\frac{|\boldsymbol{h} - \boldsymbol{h}'|}{\boldsymbol{\lambda}}\right)$$



Differential Privacy via Laplace Noise

We aim to ensure ε-differential privacy

How large should λ be?

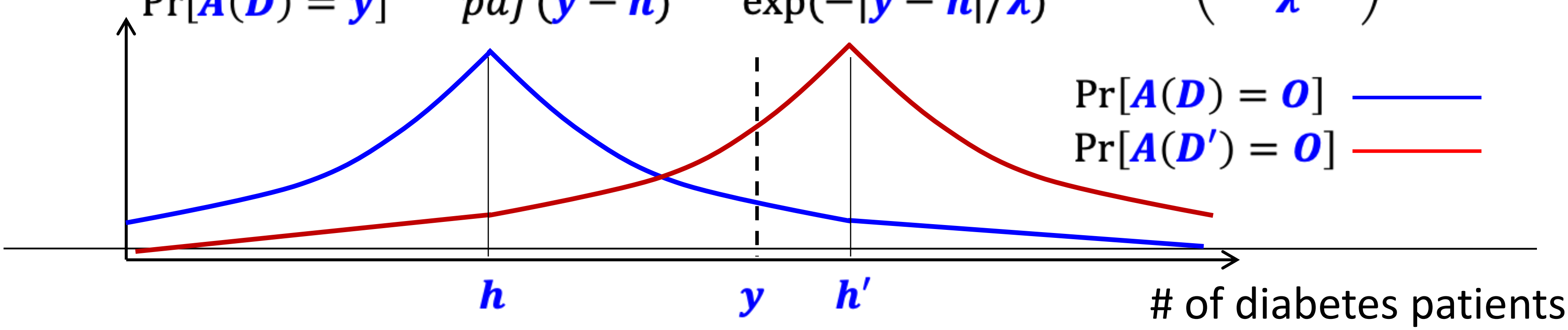
$$\exp\left(\frac{|h-h'|}{\lambda}\right) \le \exp(\varepsilon) \quad \rightarrow \quad \lambda \ge |h-h'|/\varepsilon$$

How large can $|h-h'|$ be?

Change of a patient's data would change the number of diabetes patients by at most 1, i.e., $|h-h'| \le 1$

Conclusion: Setting $\lambda \ge 1/\varepsilon$ would ensure ε-differential privacy

$$\frac{\Pr[A(D') = y]}{\Pr[A(D) = y]} = \frac{pdf(y-h')}{pdf(y-h)} = \frac{\exp(-|y-h'|/\lambda)}{\exp(-|y-h|/\lambda)} \le \exp\left(\frac{|h-h'|}{\lambda}\right)$$



Differential Privacy via Laplacian Noise

- In general, if we want to release a value $\boldsymbol{v}$

  Add Laplace noise into $\boldsymbol{v}$
- To decide the scale $\boldsymbol{\lambda}$ of Laplacian noise
- Look at the maixmum change that can occur in $\boldsymbol{v}$ (when we change one tuple in the dataset)
- Set $\boldsymbol{\lambda}$ to be proportional to the maximum change

Differential Privacy via Laplace Noise

- What if we have multiple values?
  - Add Laplace noise to each value
- How do we decide the noise scale?
  - Look at the *total change* that can occur in the values when we modify one tuple in the data
  - Total change: sum of the absolute change in each value (i.e., differences in L1 norm)
  - Set the scale of the noise to be proportional to the maximum total change
- The maximum total change is referred to as the *sensitivity* of the values
- Theorem [Dwork et al. 2006]: Adding Laplace noise of scale λ to each value ensures ε-differential privacy, if:

$$\lambda \geq (\text{the sensitivity of the values})/\varepsilon$$

Sensitivity of Queries

Histogram:

Sensitivity of the bin counts: 2

Reason: When we modify a tuple in the dataset, at most two bin counts would change; furthermore, each bin count would change by at most 1

Scale of Laplace noise required: $2/\varepsilon$

For more complex queries, the derivation of sensitivity can be much more complicated

Example: Parameters of a logistic model

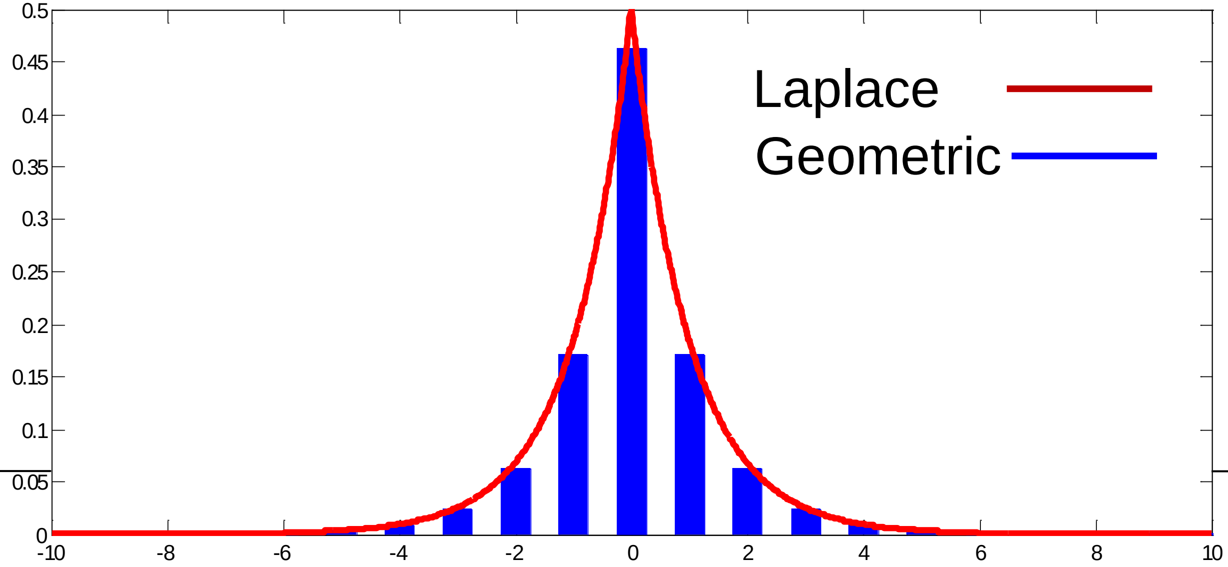
Geometric Mechanism

Suitable for queries with integer outputs [Ghosh et al. 2009]

Adds noise to each query; the noise follows a two-sided geometric distribution:

$$\Pr[x] = \frac{1-\exp(-1/\lambda)}{1+\exp(-1/\lambda)} \exp\left(-\frac{|x|}{\lambda}\right), \text{ for integer } x$$

Queries Q with sensitivity $S(Q)$ require geometric noise with: $\lambda \geq S(Q)/\varepsilon$



Exponential Mechanism

Suitable for queries with non-numeric outputs [McSherry and Talwar 2007]

Example: Top-1 frequent itemset from a transaction database ***D***

Basic Idea:

- Sample one itemset from the set of all possible itemsets
- Itemsets with higher frequencies are more likely to be sampled
- Any change of a single transaction in ***D*** should lead to only bounded change in the sampling probability

Exponential Mechanism (cont.)

Details:

Denote the frequency an itemset $\boldsymbol{I}$ as $\boldsymbol{f}(\boldsymbol{I})$

Sampling probability of $\boldsymbol{I}$: proportional to $\exp(\boldsymbol{f}(\boldsymbol{I})/\boldsymbol{\lambda})$

Why this ensures $\boldsymbol{\varepsilon}$-differential privacy:

$$\Pr[\boldsymbol{I} \text{ is sampled}] = \frac{\exp(\boldsymbol{f}(\boldsymbol{I})/\boldsymbol{\lambda})}{\sum_{\forall \boldsymbol{I}'} \exp(\boldsymbol{f}(\boldsymbol{I}')/\boldsymbol{\lambda})}$$

When change one transaction in the dataset

$\boldsymbol{f}(\boldsymbol{I})$ changes by at most 1

$\exp(\boldsymbol{f}(\boldsymbol{I})/\boldsymbol{\lambda})$ changes by a factor of at most $\exp(1/\boldsymbol{\lambda})$

For any $\boldsymbol{I}'$, $\boldsymbol{f}(\boldsymbol{I}')$ changes by at most 1

$\sum_{\forall \boldsymbol{I}'} \exp(\boldsymbol{f}(\boldsymbol{I}')/\boldsymbol{\lambda})$ changes by a factor of at most $\exp(1/\boldsymbol{\lambda})$

Thus, $\Pr[\boldsymbol{I} \text{ is sampled}]$ changes by a factor of at most $\exp(2/\boldsymbol{\lambda})$

We can achieve $\boldsymbol{\varepsilon}$-differential privacy by setting $\boldsymbol{\lambda} \geq 2/\boldsymbol{\varepsilon}$

Exponential Mechanism (cont.)

General case:

A dataset $\boldsymbol{D}$,

An output space $\boldsymbol{E}$,

A score function $\boldsymbol{f}$, such that $\boldsymbol{f}(\boldsymbol{D}, \boldsymbol{e})$ measures the "goodness" of $\boldsymbol{e} \in \boldsymbol{E}$ given dataset $\boldsymbol{D}$,

Sample any $\boldsymbol{e} \in \boldsymbol{E}$ with probability proportional to $\exp(\boldsymbol{f}(\boldsymbol{D}, \boldsymbol{e})/\boldsymbol{\lambda})$.

Theorem [McSherry and Talwar 2007]:

Achieve $\boldsymbol{\varepsilon}$-differential privacy by setting $\boldsymbol{\lambda} \geq 2\boldsymbol{S}(\boldsymbol{f})/\boldsymbol{\varepsilon}$, where $\boldsymbol{S}(\boldsymbol{f})$ denotes the sensitivity of the score function $\boldsymbol{f}$.

Composition of Differential Privacy

What if we want to compute the top-k frequent itemsets with ε-differential privacy?

Solution: Apply the previous exponential mechanism k times, each with (ε/k)-differential privacy.

Corollary from [McSherry and Tulwar 2008]:

The sequential application of m algorithms $A_1, A_2, \cdots, A_m$, each giving ε_i-differential privacy, would ensure $(\sum_{i=1}^{m} \varepsilon_i)$-differential privacy.

Variants of Differential Privacy

Alternative definition of neighboring dataset:

Two datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$, such that $\boldsymbol{D'}$ is obtained by adding/deleting one tuple in $\boldsymbol{D}$:

$$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

Even if a tuple is added to or removed from the dataset, the output distribution of the algorithm is roughly the same:i.e., the output of the algorithm does not reveal the presence of a tuple.

Refer to this version as “unbounded” differential privacy, and the previous version as “bounded” differential privacy

Variants of Differential Privacy

- **Bounded**: $\boldsymbol{D'}$ is obtained by changing the values of one tuple in $\boldsymbol{D}$.
- **Unbounded**: $\boldsymbol{D'}$ is obtained by adding/removing one tuple in $\boldsymbol{D}$.
- **Observation 1**

  Change of a tuple can be regarded as removing a tuple from the dataset and then inserting a new one

  Indication: Unbounded $\boldsymbol{\varepsilon}$-differential privacy implies bounded $(2\boldsymbol{\varepsilon})$-differential privacy.

  Proof: $\Pr[\boldsymbol{A}(\boldsymbol{D}_1) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}_2) = \boldsymbol{O}]$

  $\leq \exp(\boldsymbol{\varepsilon}) \cdot \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D}_3) = \boldsymbol{O}]$

Variants of Differential Privacy

- **Bounded**: $\boldsymbol{D'}$ is obtained by changing the values of one tuple in $\boldsymbol{D}$
- **Unbounded**: $\boldsymbol{D'}$ is obtained by adding/removing one tuple in $\boldsymbol{D}$
- **Observation 2**

  Bounded differential privacy allows us to directly publish the number of tuples in the dataset:

  $$\Pr[\boldsymbol{A}(\boldsymbol{D}) = \boldsymbol{O}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) = \boldsymbol{O}]$$

  Unbounded differential privacy does not allow this.

Variants of Differential Privacy

$(\boldsymbol{\varepsilon}, \boldsymbol{\delta})$-differential privacy:

Allows a small probability of failure

For any two neighboring datasets $\boldsymbol{D}$ and $\boldsymbol{D'}$, and for any set $\boldsymbol{S}$ of outputs:

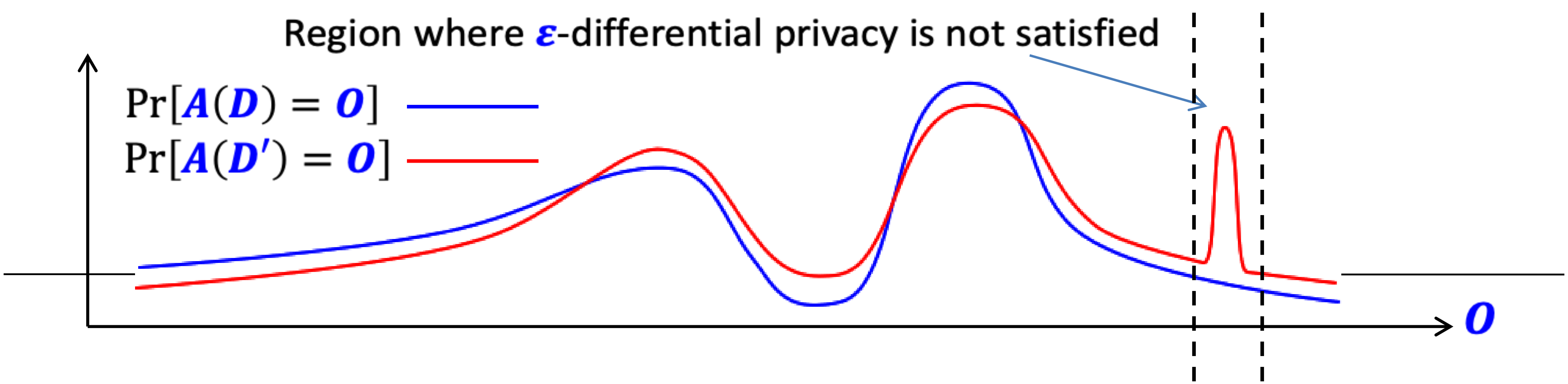
$$\Pr[\boldsymbol{A}(\boldsymbol{D}) \in \boldsymbol{S}] \leq \exp(\boldsymbol{\varepsilon}) \cdot \Pr[\boldsymbol{A}(\boldsymbol{D'}) \in \boldsymbol{S}] + \boldsymbol{\delta}$$

Relaxation of privacy provides more room for utility

Example: Use of Gaussian noise instead of Laplace noise

Composition: a set of $(\boldsymbol{\varepsilon_i}, \boldsymbol{\delta_i})$-differentially private algorithms

→ $(\sum \boldsymbol{\varepsilon_i}, \sum \boldsymbol{\delta_i})$-differential privacy



Limitations of Differential Privacy

Differential privacy tends to be less effective when there exist correlations among the tuples.

Example (from [Kifer and Machanavajjhala 2011]):

- Bob's family includes 10 people, and all of them are in a database
- There is a highly contagious disease, such that if one family member contracts the disease, then the whole family will be infected
- Differential privacy would underestimate the risk of disclosure

Summary: Amount of noise needed depends on the correlations among the tuples, which is not captured by differential privacy.