Privacy

Machine Learning Theory (MLT)

Edinburgh

Rik Sarkar

Comments on coursework

- The objective is for you to learn to read and write formal symbolic statements in ML
- The common issue in cutting edge R & D is that you have an idea/thought, but now you need to test/prove/describe it
  - And speaking vaguely does not suffice
  - We need to be precise about what we mean so that we can check it via proofs, experiments, discussions etc
- E.g. We want to say “Learning with complex models takes more data”
  - But what do we mean by Learning, complex, models, and “more”?
  - Unless we are careful about these, we will forever argue in circles over, e.g. what leanring means, and what complex means
  - Therefore, we defined PAC learning, VC dimension, $\mathcal{H}$, and the theorem of statistical learning

- Defining mathematical problems/properties
  - Often more restricted than we would like
  - But that is ok! The same exact ideas do not hold for all situations.
  - It is better that we are clear about what we mean and where it works

- In your coursework
  - Try to write the result as a property/theorem etc where you are clear about the notations, assumptions, and state it unambiguously. (e.g. see tutorial, book…)
  - Be clear about your interpretation and you will be fine (e.g. does k layers include input/layers?)
    - As long as it is not obviously wrong or inconsistent
    - Don’t think of having to answer for marking. Ask: Will this make sense to someone else reading it?
  - Keep them short and concise.
    - Accuracy is important. If it is getting into long textual arguments, reconsider.

Today

- Privacy
- Differential privacy

Privacy in data mining and machine learning

- “Big data” allows us to do ML and data mining
- Problem: Most important Datasets consist of data from people
  - Health data
  - Behaviour data: tastes in food, movies, app use, posts,
  - Financial data
- Leaking sensitive information can have adverse consequences
  - Criminal activity
  - Insurance premiums
  - Exploitative advertising and manipulation
    - Product advertising, filter bubbles, political manipulation

Often used techniques: Anonymization and pseudonymization

- Anonymization:
  - Remove name and any real world id
- Pseudonymization
  - Replace real id with a consistent random id

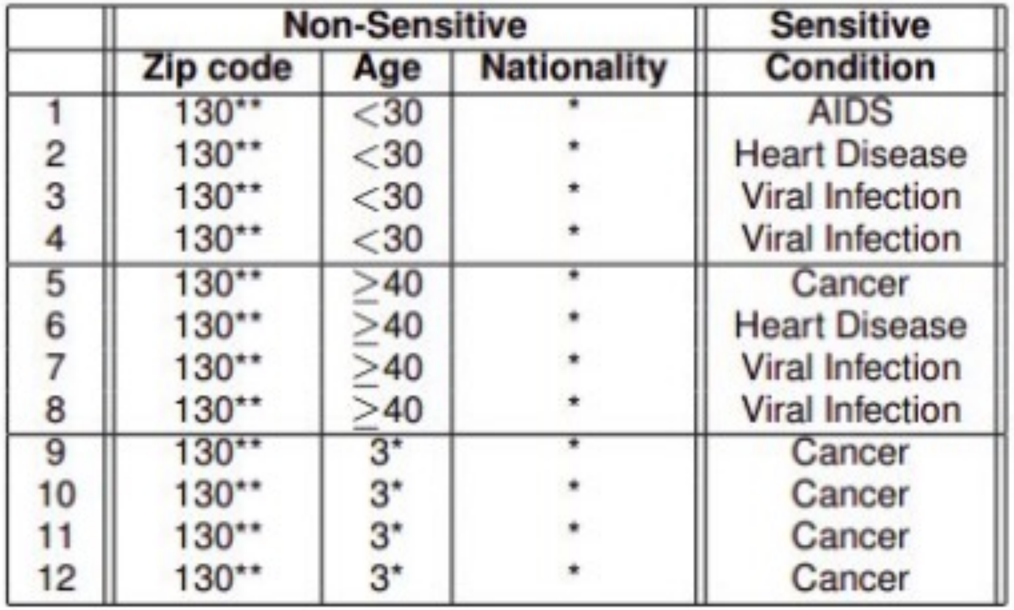
- Problems:
  - Linkage attacks: An adversary with a little bit of knowledge (e.g. age, gender, postcode) can match to the database and identify people
  - E.g. they may know someone personally, or they may have access to other datasets with that info.
  - Sometimes, information can be exploited without real-world identification (e.g. tracking cookies on web sites)

Examples

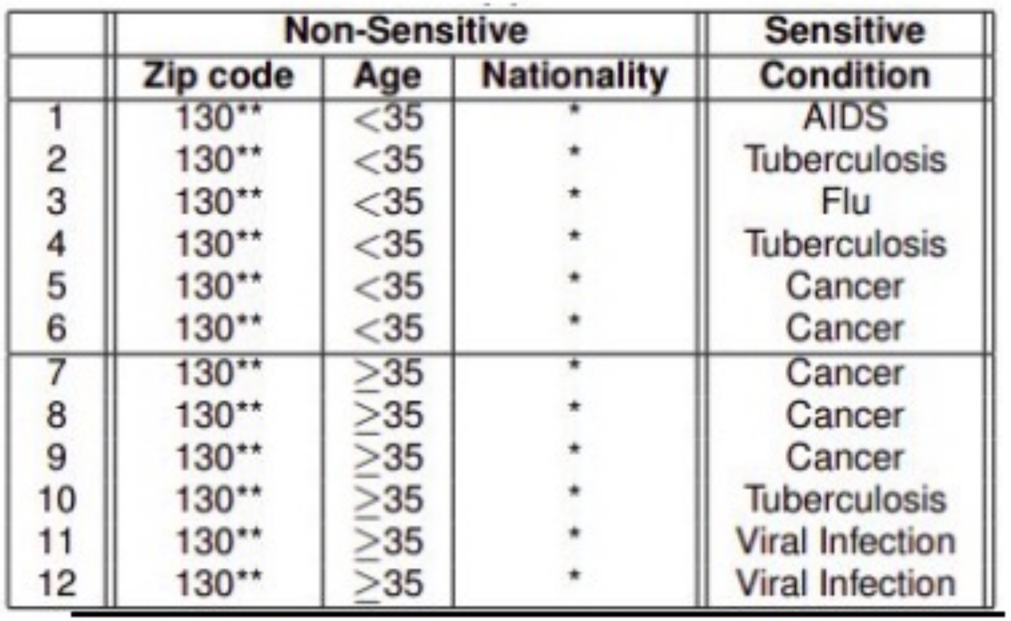
- NYC taxicab dataset
  - NYC taxi trips data released (pseudonymised) on a freedom of information request
  - Specific taxi cabs identifiable. Incomes leaked
- Medical record of Governor of Massachusetts identified by cross linking voter data
- Netflix prize data (k-anonymized) cross linked with public imdb rating

K-anonymity

- Replace an attribute by ranges of values (or representative values)
    - In groups of at least k
- Example: 4 and 6 – anonymized tables



	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<30	*	AIDS
2	130**	<30	*	Heart Disease
3	130**	<30	*	Viral Infection
4	130**	<30	*	Viral Infection
5	130**	≥40	*	Cancer
6	130**	≥40	*	Heart Disease
7	130**	≥40	*	Viral Infection
8	130**	≥40	*	Viral Infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer



	Non-Sensitive			Sensitive
	Zip code	Age	Nationality	Condition
1	130**	<35	*	AIDS
2	130**	<35	*	Tuberculosis
3	130**	<35	*	Flu
4	130**	<35	*	Tuberculosis
5	130**	<35	*	Cancer
6	130**	<35	*	Cancer
7	130**	≥35	*	Cancer
8	130**	≥35	*	Cancer
9	130**	≥35	*	Cancer
10	130**	≥35	*	Tuberculosis
11	130**	≥35	*	Viral Infection
12	130**	≥35	*	Viral Infection

K-anonymity

- Protect sensitive info: E.g. health condition
- Remove ids: E.g. voter id, name
- Group pseudo-ids: (e.g., postcode, gender, age)
  - Such that for any one person, there are at least k-1 others with identical psudo-identifiers
- Idea: Being in a group gives you better protection:
  - Adversary may be less certain which health condition applies to you.
- Still sensitive to linkage type attacks
  - E.g. if everyone in a group has the same condition, then there is no hiding

Privacy vs security

- Security: prevent unauthorized users from reading data
- Privacy: prevent authorized users from learning more than they should
  - Data should still be useful for intended tasks

Aggregate functions

- Suppose we did not release the dataset with individual values. But released only an aggregate value
  - e.g. average, sum, high-d average, count…
  - Generally: A function $f(D)$ computed on a dataset D
- Does this protect us?
- Differencing attack:
  - Datasets with and without an individual. E.g. average before and after employee joins
- Membership inference attacks
  - Find if a known person was present in the dataset
- Data reconstruction attack
  - Deduce part of the dataset from aggregate

Puzzle

What about machine learning models?

- Also a form of aggregate function (vector) computation
  - $\boldsymbol{w} = f(D)$
  - Small changes in D cause small changes in $\boldsymbol{w}$ which may reveal information about what is in the data
- Susceptible to
  - Differencing attacks
  - Membership inference
  - Data reconstruction attacks

Question: What is Privacy in your opinion?

- When would you say an aggregate value or model is private for a person in the training data?

Privacy Idea:

- The computation output should be the same irrespective of if I am there or not ie. Computation of Algorithm A is such that
  - The output $A(D)$ does not change if we add or remove x from D
  - Then there is no privacy leak
  - Adversary does not learn anything from algorithm output

Observations

- Privacy is not easy
  - Data points influence the output. Thus, some information about data points can be deduced from output (information leak)

- What the adversary can learn may depend on what the adversary knows beforehand or outside this system
  - E.g. linking other datasets

Knowledge

- Knowledge can be tricky
    - Someone has heard that you are unusually tall; they can guess that the entry with 6’6” height is likely you
        - (or, if the average in a small group is high, you were likely in the group)
    - If you are 6’, they can make a guess, but be less sure of it (lower probability)
    - If you are 5’6’, they are even less sure (there are many ~ 5’6” people, hard to be sure)
- Knowledge can be seen as probabilistic:
    - Looking at the data, you can make a better guess than before

Privacy idea 2:

- The Output $A(D)$ does not reveal much more probability about x if x is in the dataset, compared to if x is not

- E.g. If the question is membership, looking at the output (e.g. average height), it is hard to be sure that data of x was used

Differential privacy

- Idea:
- Instead of publishing the true value $f(D)$, publish a noisy version:
    - $f(D) + y$
    - Where y is a random number from a suitable distribution
    - We can write this as a simple algorithm: $A(D) = f(D) + y$

- Now an adversary cannot be sure of what the true value was
- Differential privacy lets us quantify how much privacy A ensures
- Or conversely, how much information A leaks

Differential privacy

- Idea: The output $w = A(D)$ is published.
  - It should not make it much easier to know whether x was in D or not

- Suppose D (containing x), and $D' = D \setminus \{x\}$ are two possible inputs
  - Called neighboring datasets

- We want the probability of seeing w to be similar for the two inputs:
  - $\Pr[w = A(D)] \approx \Pr[w = A(D')]$
  - Or, $\frac{\Pr[w=A(D)]}{\Pr[w=A(D')]} \approx 1$

- We want the probability of seeing w to be similar for the two inputs:
  - $\Pr[w = A(D)] \approx \Pr[w = A(D')]$
  - Or, $\frac{\Pr[w=A(D)]}{\Pr[w=A(D\prime)]} \approx 1$
  - Or, $\frac{1}{\mathrm{C}} \leq \frac{\Pr[w=A(D)]}{\Pr[w=A(D')]} \leq C$
  - For a small C just greater than 1

Differential privacy Definition

- For any neighboring Databases D and D' , A randomized algorithm A is ϵ-differentially private, if
- $\frac{1}{e^{\epsilon}} \leq \frac{\Pr[w=A(D)]}{\Pr[w=A(D')]} \leq e^{\epsilon}$
- C is written as e^{ϵ} for easier notations

Some notes

- A randomized algorithm A is ϵ-differentially private, if For any neighboring Databases D and D'
- $\frac{1}{e^{\epsilon}} \leq \frac{\Pr[w=A(D)]}{\Pr[w=A(D')]} \leq e^{\epsilon}$

- Smaller ϵ implies greater privacy
  - Zero is perfect privacy
  - Something like $\epsilon = 0.1$ is considered good privacy
- Think of ϵ representing the privacy loss due to publishing the output of algorithm A

- Also written as
    - Databases D and D' that differ in presence of one element are called neighboring databases
        - (Often without specifying which contains the element)

- For any neighboring Databases D and D' , A randomized algorithm A is ϵ-differentially private, if
    - $\frac{\Pr[w=A(D)]}{\Pr[w=A(D')]} \leq e^{\epsilon}$

The subset version

- Instead of output being a specific number w, we may want to talk about ranges, like average height between 5’5” and 5’7”
    - More useful in real valued numbers

- The corresponding subset definition follows the same template:
- For any neighboring Databases D and D', A randomized algorithm A is ϵ-differentially private, if
    - $\frac{\Pr[A(D) \in S]}{\Pr[A(D') \in S]} \leq e^{\epsilon}$

Simple Example

- Suppose there was a meeting
- We want to publish the count of the number of people in the meeting
- The published count must be ϵ-differentially private
    - That is, looking at the count, it should not be easy to deduce if a particular person was in the meeting
    - E.g. if all people in an office attended the meeting, then publishing the exact count reveals that all people attended the meeting

Solution

- Find the real count n,
- Publish $n + y$ where y is noise (a random number)

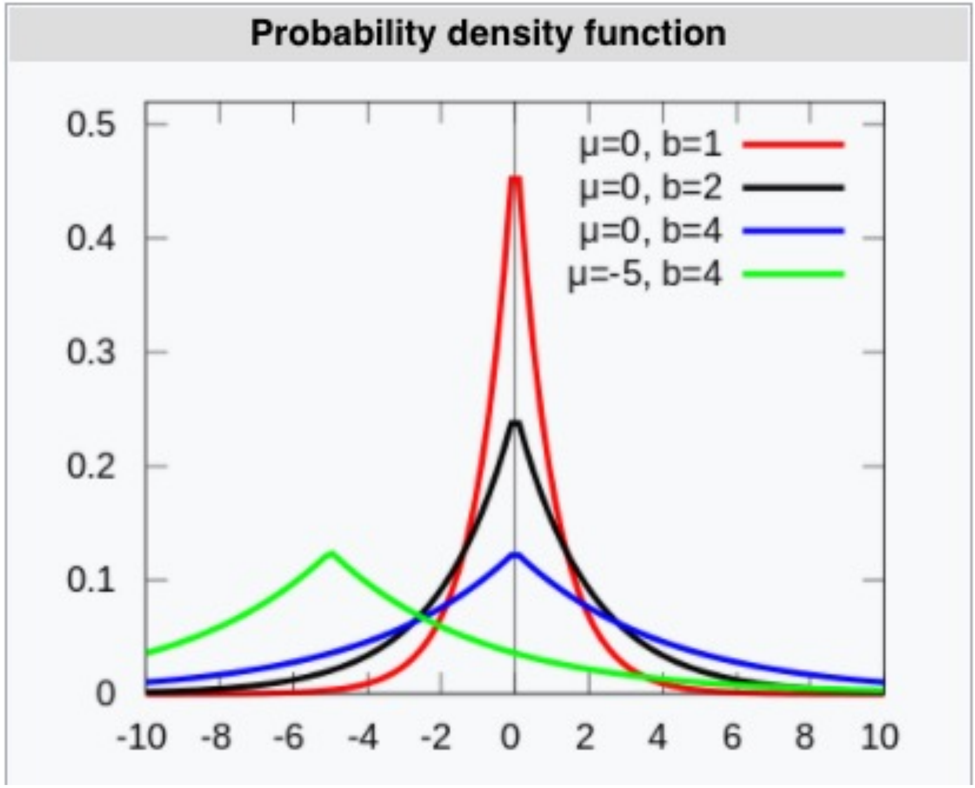
- Choice of y is important

Solution

- Find the real count n,
- Publish $n + y$ where y is noise (a random number)

- Use the Laplace distribution to get y:
  - $\Pr[y] = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$
  - μ is the mean, $2b^2$ is variance

- We will write $Lap(b)$ to mean Laplace distribution with mean 0 and variance $2b^2$



Solution

- Find the real count n,
- Publish $n + y$
  - where y $\sim Lap\left(\frac{1}{\epsilon}\right)$

- Theorem :
  - The published count is ϵ-differentially private

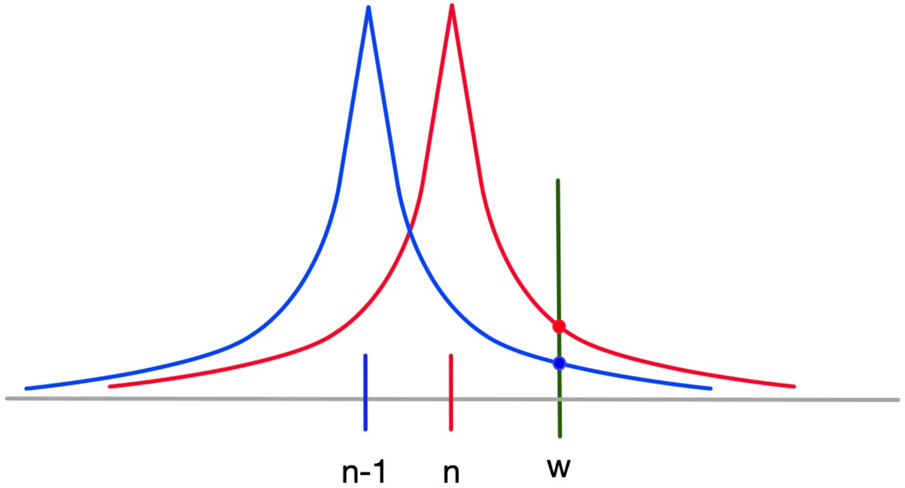
- Observe:
  - Larger variance in noise gives greater privacy.

Question:

- Instead of Laplace, what if we use a uniform random noise?
  - E.g. Suppose we output $A(n) = n + y$ where $y \in [0,5]$ uniformly

- Does this give DP? Are the output probabilities similar for all possible outputs?

- Find the real count n,
- Publish $w = n + y$
  - where y $\sim Lap\left(\frac{1}{\epsilon}\right)$
- Theorem :
  - The published count is ϵ-differentially private

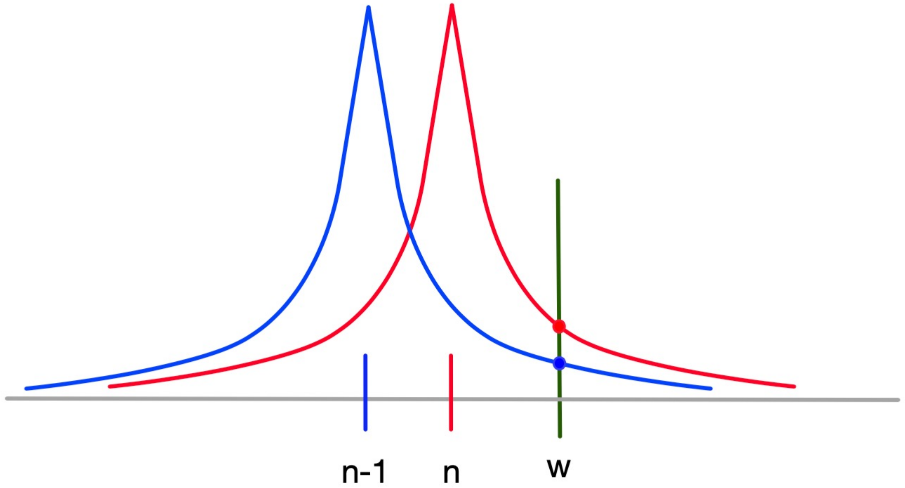


$$\Pr[y] = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$$

$$\mu = 0, b = \frac{1}{\epsilon}$$

- Proof (sketch): Any published w can occur in two ways:
- $n + y$
- $(n-1) + (y+1)$
- So it is a question of probability of y being chosen as opposed to $y + 1$
- The ratio $\frac{\Pr[noise=y]}{\Pr[noise=y+1]}$ =?

- Find the real count n,
- Publish $w = n + y$
  - where y $\sim Lap\left(\frac{1}{\epsilon}\right)$
- Theorem :
  - The published count is ϵ-differentially private



$$\Pr[y] = \frac{1}{2b} e^{-\frac{|y-\mu|}{b}}$$

$$\mu = 0, b = \frac{1}{\epsilon}$$

- Proof (sketch): Any published w can occur in two ways:
- $n + y$
- $(n - 1) + (y + 1)$

- So it is a question of probability of y being chosen as opposed to $y + 1$

- The ratio $\frac{\Pr[noise=y]}{\Pr[noise=y+1]} = \frac{e^{-|y|\epsilon}}{e^{-|y+1|\epsilon}} \leq e^{\epsilon}$

Observations

- The noise distribution needs infinite support
  - Therefore, Laplace, gaussian etc are common
  - Needs high concentration somewhere
- Concentration at zero is and useful.
  - Why do you think that is?

DP meaning

- Looking at the output $A(D)$, someone (e.g.) an adversary cannot be sure if D or D' was used in the computation
  - Thus, cannot be sure if any particular element was used in the computation
    - Since the output is equally likely in the two cases
  - Thus, protects the privacy of individual elements

- The presence of D' is hypothetical. In reality there is only one database D
  - The idea is that results are almost the same for D and D', which ensures privacy

Interpretation of differential privacy

- Suppose the adversary has powerful binoculars to see who was in the meeting room
- Or suppose, adversary knows personal information that an employee was away in a different city

- Does DP apply? What does it mean?

Differential privacy is about the computation, not the reality

- Irrespective of who actually attended the meeting, we can ask the question of who was counted in the database
- Differential privacy is about which data point was used or not used in the computation, not about what actually happened
    - Think of it as “was this row of the database actually used in the computation?”
    - Since we can never be sure of what actually happened!

The adversary who is trying to learn private information

- What do they already know?
  - They may know private information
  - They may know exactly values a database has E.g. your age, height, weight etc

- What does DP achieve?

- Remember, DP is about which data points were used in computation
- Even if adversary knows everything, DP is a probability guarantee that they cannot be sure if the data point was used in the computation!
- Since we cannot be sure of what the adversary knows in real life, DP tries to guarantee that even if they know the actual value itself, they cannot be sure if the point was actually used in computation
  - Computation is the only thing we can control!
  - Assumption of extremely strong adversary
  - In reality, adversaries are weaker, but hard to characterize that…

Adversary

- Knows the algorithm and the distribution
  - Usual assumption in security and privacy: Any design decision is known

- Does not know the noise y – the random number drawn by computer from the distribution at the time .

- Potentially knows everything about the data

Hiding Presence of elements and values of elements

- We have said that if D and D' differ in the presence of x, DP ensures that the adversary cannot be sure of which database was used
  - Thus, it hides the presence of x in the computation.
- In practice, we may care less about hiding the presence of a person, and more about hiding the value associated with the person.

- We see next, hiding the presence actually gives us a guarantee on hiding the value as well.

Hiding values

- Suppose:
  - D contains x
  - D' does not contain x
  - D'' does not contain x, and instead contains x' (where x' is an arbitrary different element)

- If Algorithm A satisfies ϵ-DP,
  - Then it provides ϵ-DP in the distinction between D and D'
  - And ϵ-DP in the distinction between D' and D''
  - Thus, it provides 2ϵ-DP in the distinction between D and D''
  - (try to write the proof/calculation for yourself)

- Thus, there is a 2ϵ-DP in x being substituted with any arbitrary x'
- The “removal version” and “replacement version” are both used frequently

- “Hiding the presence” of an element makes sense theoretically
  - It is closer to the concept that presence of it should not make much difference
- In practice, hiding the “value” may be significant
- As we see, one implies the other to a factor of 2.
- In literature, both get used, often without sufficient clarity.
  - And written as ϵ-DP
- Try to be clear and consistent – don’t confuse between the two

Hiding values: meaning

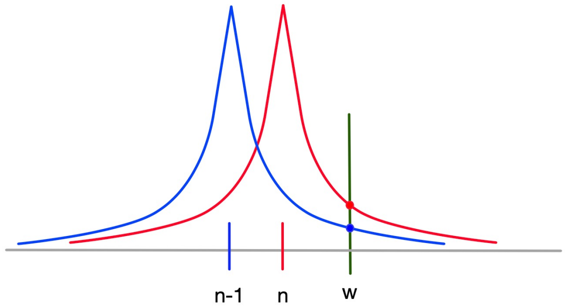
- If the adversary did not know the value, they cannot be sure of what it is
- Alternatively, even if they know the value, they cannot be sure of exactly which value was used
    - Imagine the adversary is assigning a probability estimate of what value was used
    - This probability estimate does not get better by seeing the output

Hiding values: meaning

- What does it mean to say someone “knows” something?
  - That they have a high confidence (assigns high probability at something)
  - E.g.
    - For a tall person, it may be a probability distribution with high concentration around 6’
    - For a person frequently going to a doctor, high probability of having a chronic illness, and possibly a distribution over illnesses based on more data
  - Other types of guesses based on other types of data

- Point: knowledge is probabilistic. Better knowledge means more probability/confidence at the right answer

- DP idea:
  - Seeing the DP output does not change this probability/confidence by much



What about privacy other than one person?

- What about a group of k?
- Show that an ϵ-DP algorithm is $k\epsilon$-DP for any group of size k