Differential Privacy: The Definition

Personal Notes

Ari Biswas
University Of Warwick
aribiswas3@gmail.com

1 What The Fudge Is Even Differential Privacy?

Imagine the following scenario. The local government of Wolvercote, a small village in Oxfordshire, wants to gauge public opinion on sensitive topics such as education, taxes and healthcare. So they construct a survey of questions that they want residents to fill out. The questions in the survey are of a sensitive nature, such as, *"Should we increase taxes for people in a higher income bracket?", "Should we make Covid vaccines mandatory?", "Should education be free for all and the costs come from public taxes?"*

The residents are worried, that if their opinions were to be made public, they might be subject to the public backlash and their village life will end up being socially divided (as is warrant these days on the internet). Thus, the government has promised its residents that their opinion will not be leaked at any cost. Only then are the residents willing to participate in this survey.

Now the government do not want collect all this information from the residents just to let data sit in a secure vault. They obviously want to compute statistics on this data that encapsulates public opinion. Furthermore, they plan to *publicly release* these statistics to justify any future policy changes. For example, they might want to know which policy on the questionnaire received the most votes, such as what fraction of its residents wish for mandatory vaccinations for Covid. Let $X = (x_1, \dots, x_n)$ denote the survey responses of the residents and $q_i(x_i) \in \{0, 1\}$ denote person i's response. The government computes $f(X) = \frac{1}{n} \sum_{i \in [n]} q_i(x_i)$ and releases for everyone to see.

They do this every year to justify that the political changes are in accordance with the majority of the populations interests. Now imagine, a new person moves into the village. Let $X' = (x_1, \dots, x_n, x_{n+1})$ denote the survey responses and $f(X') = \frac{1}{n+1} \sum_{i \in [n+1]} q_i(x_i)$ denote the new results with the new person included. Also imagine that the residents have not changed their opinions from the previous year. They stick by their choices (A reasonable assumption given these are sensitive topics on which people will unlikely change their mind). $f(X')$ and $f(X')$ are publicly available for everyone to see. However, now anyone subtracts this years results from last years results, the private opinion of this new person is compromised.

This is undesirable. The issue here is that the statistic that is being released is computed **deterministically** as a function of resident opinions (and everyone knows what the function is – in the above example, it's just the sum of inputs). As long as the outputs are such deterministic functions, there is always a chance that some information about the inputs will be leaked.

So what do we do to prevent some leakage ? We use randomness. That is to say, we make the final answer random. Of course, you might say, if the government makes decisions on outputs generated uniformly at random, the whole survey is pointless (albeit, there will 0 leakage about any inputs from the output) So clearly, we cannot use uniformly generated randomness. So what randomness can we use? The goal of Differential Privacy (DP) is to formally answer the above questions. That is, how much randomness should be used and what is the effect of using such randomness.

2 DP: The Definition

This section explores the motivation behind the definition of differential privacy, and is based on [Vad13], Chapter 3 of [DR+14], and Chapter 1 of [Vad17]. The following informal statement summarises the idea of differential privacy (DP).

> We have some public function f that takes more n inputs $X = (x_1, \ldots, x_n)$ and produces a single output y. If the output in its exact form could reveal bits of information about the inputs that were used to compute said output. Thus, instead, instead of outputting the exact output, we output a random approximate output. The randomness ensures that we do not reveal too much information about any of the inputs, even in the worst case, to a powerful adversary. However, we also do not want to so much randomness such that the new output is useless.

That's the story. We spend the remainder of this section formalising what "leaking too much information", "worst case", "powerful adversary" and "useless/useful approximate output" mean.

General Assumptions For $n \in \mathbb{N}$, we will always assume that we wish to compute $f : \mathcal{X}^n \to \mathcal{Y}$ for any two (possibly infinite or uncountable) sets $\mathcal{X}$ and $\mathcal{Y}$. We will also assume that we do not need the exact answer for $y = f(x_1, \ldots, x_n)$[1]. We also assume that a trusted functionality exists, so we can send input X to this functionality and compute f (or a randomised version of f) for us[2]. By a *powerful adversary* A we will always mean a Turing machine with unlimited memory and compute. The *worst case* refers to the situation where A knows all but one of the inputs in X. Next, we define what we mean by a mechanism.

> **Definition 1** (Mechanism) Fix $n \in \mathbb{N}$ and let $\mathcal{Q}$ be a family of functions[a] such that for any $f \in \mathcal{Q}$, we have $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are sets. The only assumption about $\mathcal{X}$ and $\mathcal{Y}$ we will make is that there exists some base measure on $\mathcal{Y}$. A mechanism is defined by a function $\mathsf{M} : \mathcal{X}^n \times \mathcal{Q} \to \Delta(\mathcal{Y})$, where $\Delta(\mathcal{Y})$ is a probability distribution over $\mathcal{Y}$. On receiving an input $X \in \mathcal{X}^n$ and a function $f \in \mathcal{Q}$, the mechanism's output is a sample $z \xleftarrow{\$} \mathsf{M}(X, f)$.
>
> [a] We will often use the terms function and query interchangeably.

We say a mechanism M is ε-differentially private if the conditions in Definition 2 hold:

> **Definition 2** (Information Theoretic Pure DP) Fix $\kappa \in \mathbb{N}$ as the security parameter. Let $n = \mathtt{poly}(\kappa)$ and $\varepsilon \geq 0$. Let $\mathcal{Q} = \{f : \mathcal{X}^n \to \mathcal{Y}\}$ denote a family of functions. Further, assume that there is a probability measure on $\mathcal{Y}$. A mechanism $\mathsf{M} : \mathcal{X}^n \times \mathcal{Q} \to \Delta(\mathcal{Y})$ satisfies (ε, δ) differential privacy if for every two neighbouring datasets X and X' such that X and X' differ by only one element and for every query $f \in \mathcal{Q}$ we have for all $T \subseteq \mathcal{Y}$
>
> $$\Pr_{Y \xleftarrow{\$} \mathsf{M}(X,f)}[Y \in T] \leq e^{\epsilon} \Pr_{Y \xleftarrow{\$} \mathsf{M}(X',f)}[Y \in T] \tag{1}$$

[1] Note if we did, then there is nothing we can do about the information y leaks. We can only prevent leakage by not releasing y in the clear.

[2] This assumption abstracts any issues that may arise while computing f. We will not focus on how things are computed, but more so, if it was computed as prescribed, what are the consequences.

Another way to state the above definition is say that the the max divergence of $\mathsf{M}(X, f)$ and $\mathsf{M}(X', f)$ is upper bounded by ε i.e. the two distributions are close (even in the worst case).

Definition 3 (Max Divergence) For two distributions $\mathcal{D}_A, \mathcal{D}_B \in \Delta(\mathcal{Y})$, max-divergence is defined by

$$D_\infty(A||B) = \max_{T \subseteq \mathcal{Y}} \log\left(\frac{\Pr[A \in T]}{\Pr[B \in T]}\right) \tag{2}$$

Remark 1 The for all criterion in Definition 2 is the same as the $\max_{T \subseteq \mathcal{Y}}$ condition in the above definition. If the max is upper bounded, then, all subsets are also upper bounded.

How To Read The Definition Going back to our story above, for a mechanism M, the randomised output for the function f on input $X \in \mathcal{X}^n$ is an independent sample from the distribution $\mathsf{M}(X, f)$. Through the specifications of Equation (1), DP requires that on neighbouring datasets X and X', the distributions $\mathsf{M}(X, f)$ and $\mathsf{M}(X', f)$ are "roughly" the same (sameness is measured in max-divergence). The smaller the value of ε, the more the same-ness. Thus, our adversary A, despite having full knowledge of $\mathsf{M}(X', f)$, by just looking at a sample y, cannot tell if the sample came from $\mathsf{M}(X, f)$ or $\mathsf{M}(X', f)$. This implies A could not have learned too much about any input used to compute y that it did not otherwise know (otherwise, it would be able to distinguish between the distributions). We will further formalise what not learning too much or distributions being roughly the same mean. But for now, we have a grasp on what a worst case adversary is and what conditions we need for them to not learn too much information about the inputs.

Next we discuss usefuleness. Consider the mechanism $\mathsf{M}_{\texttt{Uniform}}$, which outputs a uniform distribution over $\mathcal{Y}$ as its output regardless of the inputs X. Using the definition above, such a mechanism is (0)-DP i.e. A, who knows X', learns nothing[3] about x[4] . But such a mechanism is useless for any downstream use, as the output also contains no information about $f(X)$. So, for M to be useful, we must limit how random we can make the output. That limit is described by upper bounding the utility of a mechanism:

Definition 4 (Utility) Fix $n \in \mathbb{N}$ and $\varepsilon \geq 0$. Let $\mathcal{Q} = \{f : \mathcal{X}^n \to \mathcal{Y}\}$ denote a family of functions whose output we wish to make differentially private. Further, assume that a distance metric $||\cdot||_1$ is well defined on $\mathcal{Y}$. Let M be an ε-DP mechanism for $f \in \mathcal{Q}$. For any $X = (x_1, \ldots, x_n)$, we have $y = f(X)$. Then, the utility of M is defined as

$$U(\mathsf{M}, f, X) := \mathbb{E}_{\widetilde{Y} \xleftarrow{\$} \mathsf{M}(X,f)}\left[||\widetilde{Y} - y||_1\right] \tag{3}$$

Thus, a reasonable utility condition on M, would be to say the average error between samples and the true answer is 0. The expectation of $\mathsf{M}(X, f)$ dictates the usefulness and the variance of $\mathsf{M}(X, f)$ governs privacy.

2.1 But Why This Definition?

We want $\mathsf{M}(X, f)$ and $\mathsf{M}(X', f)$ to be roughly the same. In statistical distance, we have a well-established definition of what it means for distributions to be close. So, why not use statistical distance instead?

Remark 2 (Why Not Statistical Distance?) A very natural question is why use max-divergence instead of just using the well-accepted statistical closeness definition, i.e. replace Equation (1) with Equation (4) in Definition 2, for some proximity parameter $\delta \in [0, 1]$. In other words, why not ask the definition to be

[3] Not even one bit of information.
[4] Here $x = X \setminus X'$, the input that is in X but not X'

$$d_{\text{TV}}\Big(M(X,f), M(X',f)\Big) \leq \delta \tag{4}$$

First, notice that Equation (1), implies Equation (4) in that if we have DP, then the two distributions are statistically close.

$$d_{\text{TV}}\Big(M(X,f), M(X',f)\Big) \leq \delta = 1 - e^{-\varepsilon} \leq \varepsilon$$

However, we cannot go from Equation (4) to Equation (1). If we started with Equation (4) as the definition, our mechanism would be either non-private or useless i.e. we cannot define a meaningful value for δ for which a DP mechanism is useful. Consider the following settings:

1. $\delta \leq \frac{1}{2n}$: We want output of M on neighbouring inputs to be very close.
   Let $\mathcal{X} = \{0,1\}$, consider any dataset $X \in \{0,1\}^n$ and consider a dataset $X^{''}$ that differs from X in 2 locations. For any f, by the triangle inequality we have

$$\begin{aligned} d_{\text{TV}}\Big(\mathsf{M}(X,f), M(X^{''},f)\Big) &\leq d_{\text{TV}}\Big(\mathsf{M}(X,f), M(X',f)\Big) + d_{\text{TV}}\Big(M(X',f), M(X^{''},f)\Big) && (5)\\ &\leq \frac{1}{2n} + \frac{1}{2n} && (6)\end{aligned}$$

   By an inductive argument, for any two datasets X and $\widetilde{X}$ that differ at n locations, we have

$$\begin{aligned} d_{\text{TV}}\Big(\mathsf{M}(X,f), M(\widetilde{X},f)\Big) &\leq \frac{n}{2n} && (7)\\ &= \frac{1}{2} && (8)\end{aligned}$$

   Let's fix $\widetilde{X}$ to be the all 0's dataset[5]. Ideally, for M to be useful, we want the distribution $\mathsf{M}(X,f)$ to depend on X. Otherwise, it is not useful by our definition of utility. In this situation when $\delta \leq \frac{1}{2n}$, regardless of what X is, we need $\mathsf{M}(X,f)$ to be at most at a distance of $\frac{1}{2}$ from a distribution that does not depend on X at all. In other words, at least half the time, M outputs values from $\mathcal{Y}$ independent of the inputs.

2. $\delta \geq \frac{1}{2n}$: Ok, maybe the previous assumption was too aggressive. For a given X, consider a mechanism M that outputs each input in X with probability $\frac{1}{2n}$. As there are n inputs, the probability of releasing an input in the clear is $\frac{1}{2}$. However, for a pair of neighbouring datasets of size n that differ at one location, we have $d_{\text{TV}}\Big(\mathsf{M}(X,f), M(X',f)\Big) = \frac{1}{n} > \frac{1}{2n}$. This definition that says M is private, but M outputs some input $x \in X$ in the clear with probability $\frac{1}{2}$.

[5]This could be any constant dataset, there is nothing special about the all 0's dataset.

So statistical distance did not work and using max divergence seems to be a more general claim. But we don't want an adversary to learn new information – so why not use a Bayesian definition? After all Bayesians have argued about belief propagation and updated beliefs forever. Or why not use Simulation like in Zero Knowledge? These are well-established techniques to define gaining of knowledge. Why did we end up using max divergence instead?
It turns out that it would not have mattered. All the definitions say the same.

2.2 Bayesian Interpretation Of The Definition

For brevity, in this section, for any $f \in \mathcal{Q}$, we write $\mathsf{M}(X, f)$ simply as $\mathsf{M}(X)$. For each $i \in [n]$, let $\mathcal{D}(X_i)$ denote A's prior belief of the value of $x_i \in X$, for some dataset $X \in \mathcal{X}^n$. After seeing $y \xleftarrow{\$} \mathsf{M}(X)$, let $\mathcal{D}(X_i|\mathsf{M}(X) = y)$ denote A's updated belief about the value of x_i. We will show that if M is ε-DP, then A's prior and updated belief is roughly the same (in statistical distance). Thus A could not have learned too much information about the unknown input used in the computation. We describe this phenomenon of A's belief not changing significantly despite seeing the mechanism output as Baysian DP. Formally, we state the following theorem

Theorem 1 (ε-DP implies Bayesian Privacy) Let M be an ε-DP mechanism. For $n \in \mathbb{N}$, let $\mathcal{D}$ be a joint distribution over $\mathcal{X}^n \times \mathcal{X}^n$, such that $\mathcal{D}(X \sim X') = 1$, i.e. whenever we sample from $\mathcal{D}$ we get a pair of neighbouring datasets X and X'. Let X' be a dataset that A has full knowledge of and X be a neighbouring dataset that differs in only position $i \in [n]$. Let $\mathcal{D}(X_i)$ denote the adversary A's prior belief about $x_i \in X$ and $\mathcal{D}(X_i|\mathsf{M}(X) = y)$ denote A's updated belief about the value in position i in X after seeing the output y. If M is ε-DP then

$$d_{\mathsf{TV}}(\mathcal{D}(X_i), \mathcal{D}(X_i|\mathsf{M}(X) = y)) \leq 2\varepsilon \tag{9}$$

Proof. Let $X = X' \cup \{x\}$. Let X_i be the random variable describing A's prior belief about the actual value of x.

$$\begin{aligned}
\Pr[X_i = x|M(X)] &= \frac{\Pr[X_i = x \wedge M(X_i \cup X') = y]}{\Pr[M(X)} && (10)\\
&= \frac{\Pr[M(X_i \cup X') = y|X_i = x]}{\Pr[M(X)} \Pr[X_i = x] && (11)\\
&\leq e^{\varepsilon} \frac{\Pr[M(X) = y]}{\Pr[M(X')} \Pr[X_i = x] && (12)\\
&\leq e^{\varepsilon} e^{\varepsilon} \Pr[X_i = x] && (13)\\
&= e^{2\varepsilon} \Pr[X_i = x] && (14)\\
&\leq (1 + 2\varepsilon) \Pr[X_i = x] && (15)
\end{aligned}$$

Equations (10) and (11) comes from Bayes Rule, and Equation (12) comes from the assumption that M is ε-DP i.e, $e^{-\varepsilon} \leq \frac{\Pr[M(X)]}{\Pr[M(X')]}$ and (13) comes from DP again i.e. $\frac{\Pr[M(X)]}{\Pr[M(X')]} \leq e^{\varepsilon}$. The last inequality comes from Taylor series.

□

That is good news. Our definition of DP has nice implications. Looking at just the output does not update A's belief on what inputs were used to do the computation. But what if we started with this assumption? It turns out that that converse also holds, and Bayesian Privacy implies DP.

Theorem 2 (Bayesian Privacy also implies ε-DP)

$$d_{\text{TV}}\Big[\mathcal{D}(X_i), \mathcal{D}(X_i|M(x_i \cup \widetilde{X}) = y)\Big] \leq \varepsilon \implies \mathsf{M} \text{ is } \varepsilon - \text{DP}$$

Proof. Without loss of generality, assume that[6] $\mathcal{X} = \{0, 1\}$. Let X, X' be two neighbouring datasets that differ at a single position $i \in [n]$. Let $X = (0, \widetilde{X})$ and $X' = (1, \widetilde{X})$. Let X_i be A's prior belief about input x_i in the dataset X. We have by the assumption of Bayesian privacy that

$$d_{\text{TV}}\Big[\mathcal{D}(X_i), \mathcal{D}(X_i|M(x_i \cup \widetilde{X}) = y)\Big] \leq \varepsilon \tag{16}$$

The adversary A does not know the value of the input at position i. Given X and X', X_i could only be 0 or 1. Assuming the adversary has no prior information about which one it is, we get

$$\Pr[X_i = 0] = \Pr[X_i = 1] = \frac{1}{2}$$

Thus for some $\gamma \in [0, 1]$, after observing y, A's updated beliefs are,

$$\Pr[X_i = 0|M(X_i \cup \widetilde{X}) = y] = \frac{1+\gamma}{2} \qquad \Pr[X_i = 1|M(X_i \cup \widetilde{X}) = y] = \frac{1-\gamma}{2} \tag{17}$$

By the Bayesian Privacy assumption from Equation (16) we have

$$d_{\text{TV}}\Big(\mathcal{D}(X_i = 0), \mathcal{D}(X_i = 0|M(x_i \cup \widetilde{X}) = y)\Big) = \frac{\gamma}{2} \leq \varepsilon \tag{18}$$

$$d_{\text{TV}}\Big(\mathcal{D}(X_i = 1), \mathcal{D}(X_i = 1|M(x_i \cup \widetilde{X}) = y)\Big) = \frac{\gamma}{2} \leq \varepsilon \tag{19}$$

Thus we get $\gamma \leq 2\varepsilon$.

$$\frac{\Pr[\mathsf{M}(X) = y]}{\Pr[M(X') = y]} = \frac{\Pr[M(0 \cup \widetilde{X}) = y]}{\Pr[M(1 \cup \widetilde{X}) = y]} \tag{20}$$

$$= \frac{\Pr[X_i = 0|M(x_i \cup \widetilde{X}) = y] \Pr[X_i = 0]}{\Pr[X_i = 1|M(x_i \cup \widetilde{X}) = y] \Pr[X_i = 1]} \tag{21}$$

$$= \frac{1+\gamma}{1-\gamma} \tag{22}$$

$$\leq e^{2\varepsilon} \tag{23}$$

In equation (23), we use our bayesian privacy assumption to upper bound the privacy loss (or max-divergence of $\mathsf{M}(X)$. □

[6] By making $\mathcal{X} = \{0, 1\}$, we are actually giving A as much information as possible to infer things about the input. The unknown x_i can only take two values. If it could take on more value, it just reduces the information the adversary can get.

2.3 Simulation Styled Definition

We might also ask why not define an ideal world, where an ideal adversary called simulator (Sim) has access to all but one input in X. Then to define security, show that there exists a simulator that can learn whatever a real world adversary can. More formally,

Definition 5 A PPT mechanism M is simulation differentially private if there exists a polynomial time simulator Sim such that for all $f \in \mathcal{Q}$, and any $X \in \mathcal{X}^n$, for any $i \in [n]$

$$D_\infty \left(\mathsf{M}(X, f) || \mathsf{Sim}(X', f)\right) \leq \varepsilon$$

$$D_\infty \left(\mathsf{Sim}(X', f) || \mathsf{M}(X, f)\right) \leq \varepsilon$$

where $X_{-i} = X \setminus \{x_i\}$.

It is easy to see that

Theorem 3 If M is differentially private, then M is simulation differentially private.

Proof. The simulator just computes $\mathsf{M}(X_i, f)$ for any $i \in [n]$. By the definition of DP, we get what we need. □

Theorem 4 If M is ε-simulation private, then M is 2ε-differentially private.

Proof. By Sim-DP definition, for any $T \subseteq \mathcal{Y}$, we have

$$D_\infty \left(\mathsf{M}(X, f) || \mathsf{Sim}(X', f)\right) = \frac{\Pr[\mathsf{M}(X, f) \in T]}{\Pr[\mathsf{Sim}(X', f) \in T]} \leq e^\varepsilon \tag{24}$$

$$D_\infty \left(\mathsf{Sim}(X', f) || \mathsf{M}(X, f)\right) = \frac{\Pr[\mathsf{Sim}(X', f) \in T]}{\Pr[\mathsf{M}(X, f) \in T]} \leq e^\varepsilon \tag{25}$$

Multiplying the two equations, we get

$$e^{-2\varepsilon} \leq \frac{\Pr[\mathsf{M}(X', f) \in T]}{\Pr[\mathsf{M}(X, f) \in T]} \leq e^{2\varepsilon}$$

□

References

D

[DR+14] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

V

[Vad13] Salil Vadhan. Notes on the definition of dp, 2013.

[Vad17] Salil Vadhan. The complexity of differential privacy. *Tutorials on the Foundations of Cryptography: Dedicated to Oded Goldreich*, pages 347–450, 2017.