6.885: Advanced Topics in Data Processing Fall 2013

Lecture 20: Introduction to Differential Privacy

Stephen Tu

1 Overview

This lecture aims to provide a very broad introduction to the topic of *differential privacy*. Generally speaking, differential privacy is an area of research which seeks to provide rigorous, statistical guarantees against what an adversary can infer from learning the results of some randomized algorithm. The definition was first proposed in Cynthia Dwork's ICALP paper [5]. Since then, differential privacy has become an increasingly popular area of research, with many contributions in terms of both theoretical analysis and practical instantiations.

2 Threat model

Recall the various adversarial scenarios discussed in lecture. Differential privacy addresses the case when a trusted data curator wants to release some statistic over its data without revealing information about a particular value itself.

As a concrete example, suppose a hospital has a database of patient records, each record containing a binary value indicating whether or not the patient has some form of cancer. Presumably, patients would not want others to find out whether or not they have cancer. However, say the hospital wishes to release the *total* number of patients with a particular form of cancer (a summation over these binary values) for scientific reasons. Differential privacy address the question of, given the total number of patients with cancer, whether or not an adversary can learn if a particular individual has cancer.

In this example, we have made several assumptions. First, we have assumed that the hospital can be trusted; that is, (a) the hospital's database is secure from the outside (so the adversary cannot simply hack the database to get the answer) and (b) it is in the interest of the hospital to protect individual patient privacy. Second, we have assumed (obviously) that the adversary itself does not already know the answer.

3 Previous efforts

Differential privacy is not the first framework which tries to address this question. Indeed, there have been many previous efforts which we will not discuss in more detail other than to say that they are mostly considered to be "broken" in the sense there are well known attacks. An example of this is k-anonymity [14]. The success of differential privacy stems a lot from its rigorous definition, which we will now discuss.

4 Definition

Let $\mathcal{A} : \mathcal{D}^n \to \mathcal{Y}$ be a randomized algorithm. Let $D_1, D_2 \in \mathcal{D}^n$ be two databases that differ in at most one entry (we call these databases *neighbors*). Let $\mu(\cdot|D)$ be the density of values on the distribution over $\mathcal{Y}$ induced by $\mathcal{A}(D)$, where the randomness comes from the coin tosses of $\mathcal{A}$.

Definition 1. *Let $\epsilon > 0$. Define $\mathcal{A}$ to be* ϵ-differentially private *if for all neighboring databases D_1, D_2, and for all $y \in \mathcal{Y}$, we have*

$$\frac{\mu(y|D_1)}{\mu(y|D_2)} \leq \exp(\epsilon)$$

By convention, if both values in the numerator and denominator are 0, we say the ratio is 1. It is clear from the definition that lower values of ϵ correspond to more "privacy".

Observation 2. *Because we can switch D_1 and D_2 interchangeably, Definition 1 implies that*

$$\exp(-\epsilon) \leq \frac{\mu(y|D_1)}{\mu(y|D_2)} \leq \exp(\epsilon)$$

Since $\exp(\epsilon) \approx 1 + \epsilon$ *for small ϵ, then we have roughly*

$$1 - \epsilon \lesssim \frac{\mu(y|D_1)}{\mu(y|D_2)} \lesssim 1 + \epsilon$$

Intuition. We can think of differential privacy as a game between two parties Alice and Bob. For simplicity, assume that $\mathcal{A}$ is permutation invariant (order of inputs does not matter) and the space $\mathcal{D}$ is finite (say $|\mathcal{D}| = m$). Alice picks an arbitrary $D \in \mathcal{D}^n$. Let $D_{\neg n} = (d_1, ..., d_{n-1})$, and let $D_{n,m} = (d_1, ..., d_{n-1}, d_n = m)$, where $d_n = m$ means d_n takes on the m-th value of $\mathcal{D}$. Then Alice gives Bob the tuple $(D_{\neg n}, y = \mathcal{A}(D))$. Bob must then guess correctly the value of d_n. Assuming Alice draws d_n uniformly at random, Bob's maximum likelihood guess for d_n is

$$\operatorname*{argmax}_{j \in [m]} \mu(y|D_{n,j})$$

That is, for each possible value j of d_n, Bob can learn the distribution induced by $\mathcal{A}(D_{n,j})$, and then pick the value of j which assigns highest probability to y. But if $\mathcal{A}$ satisfies ϵ-differential privacy, then we have for all $i, j \in [m]$

$$|\mu(y|D_{n,i}) - \mu(y|D_{n,j})| \lesssim \epsilon$$

In other words, (a) Bob will have very low confidence in his estimate of $d_{n,m}$ and (b) Bob will not be able to win much better than random guessing.

To make this more concrete, let us take a simple example and bound the probability of Bob willing. Suppose both $\mathcal{D} = \mathcal{Y} = \{0, 1\}$. Then Alice will pick $d_n = 0$ with probability 1/2 and $d_n = 1$ with probability 1/2. Suppose that $\mu(y = 0|D_{n,0}) > \mu(y = 0|D_{n,1})$ (the other cases are similar). Then the probability of Bob winning is given by

$$\begin{aligned}\Pr[\text{Bob wins}] &= \Pr[\text{pick } d_n = 0 \wedge d_n = 0] + \Pr[\text{pick } d_n = 1 \wedge d_n = 1] \\ &= \frac{1}{2}\Pr[\text{pick } d_n = 0 | d_n = 0] + \frac{1}{2}\Pr[\text{pick } d_n = 1 | d_n = 1] \\ &= \frac{1}{2}\mu(y = 0|D_{n,0}) + \frac{1}{2}\mu(y = 1|D_{n,1}) \\ &= \frac{1}{2}\mu(y = 0|D_{n,0}) + \frac{1}{2}(1 - \mu(y = 0|D_{n,1})) \\ &\leq \frac{\mu(y = 0|D_{n,1})}{2}(\exp(\epsilon) - 1) + \frac{1}{2} \\ &\leq \frac{\exp(\epsilon)}{2} \lesssim \frac{1}{2} + \frac{\epsilon}{2}\end{aligned}$$

Thus, given *every* data point but the n-th point plus an outcome of $\mathcal{A}$ on all of the data, Bob can barely do better than random guessing when trying to learn the remaining data point.

Note that, on the other hand, suppose we do not enforce differential privacy. Consider the function $f(D)$ given by

$$f(D) = \begin{cases} 1 & \text{if } \sum\limits_{i=1}^{n} I(d_i = 1) > \sum\limits_{i=1}^{n} I(d_i = 0) \\ 0 & \text{otherwise} \end{cases}$$

where $I(\cdot)$ is the indicator function. If we restrict ourselves to cases where $|D|$ is odd and have Alice and Bob play the game described above, then $\Pr[\text{Bob wins}] = 1$.

4.1 Statistical guarantees

Now that we have a definition of what it means to be private, it is natural to ask what statistical guarantees are provided by an algorithm which satisfies the definition. Naturally, we might hope that the information learned about an individual by the output of some algorithm is no more than the information we can learn about that individual *without* access to the output. Informally, we will call this *pure semantic privacy.* Unfortunately, external information makes such a privacy definition impossible (without destroying all utility).

As a silly contrived example, suppose we know (already) that Alice is a chain smoker. Then suppose we release the results of a study which indicates that chain smokers have a much greater chance of getting lung cancer (pretend we did not know this already). Then we have just learned that Alice is predisposed to lung cancer, even though Alice did not even participate in the study!

Given this intuition, we therefore must aim for more relaxed definitions of privacy than pure semantic privacy. In this section, we provide intuition that differential privacy achieves a *relaxed* version of semantic privacy. Informally, differential privacy states (which we saw previously in a specific example) that an adversary with access to the output of an algorithm will learn roughly the same information whether or not a single user's data was included or not. So if a user is conflicted about whether or not to participate in a statistical database, the user can be assured that her participation will not drastically affect the outcome of functions run on the database.

We now seek to formalize this argument. Suppose Alice participates in a statistical database. Mathematically, we need to bound the *statistical difference* between the posterior beliefs b_1, b_2, where b_1 is a posterior on the values in the database given an output $y = \mathcal{A}(D)$ where D includes Alice's data, and b_2 is a posterior on the database given the same output $y = \mathcal{A}(D')$, where D' does not include Alice's data.

Differential privacy implies (relaxed) semantic privacy. The development of this section is a very simplified presentation of [10]. Here, for simplicity we restrict ourselves to finite $\mathcal{D}$ and $\mathcal{Y}$. As a preliminary, let us define the *statistical difference* between two distributions X, Y on the same discrete probability space Ω as

$$SD(X,Y) = \max_{\rho \in \Omega} |X(\rho) - Y(\rho)|$$

Let us assume our randomized algorithm $\mathcal{A}$ satisfies ϵ-differential privacy, and once again that it is permutation invariant. Let $b(D)$ denote an adversary's prior belief on databases $D \in \mathcal{D}^n$, and $b(D|y)$ denote the posterior belief on databases, given an output $y \in \mathcal{Y}$. Let $D_{\neg n}$ denote a database where we keep the first $n-1$ values of D, but replace the n-th value with some arbitrary $d_n \in \mathcal{D}$. Consider an alternate world where we use a different randomized algorithm $\mathcal{A}'(D) = \mathcal{A}(D_{\neg n})$, and let $b'(D|y)$ denote the posterior belief in the alternate world.

We will now argue that for all $D \in \mathcal{D}^n$ and for all $y \in \mathcal{Y}$

$$SD(b(D|y), b'(D|y)) \leq \exp(2\epsilon) - 1$$

Intuitively, this means that the adversary's posterior belief, upon seeing an output y, is insensitive to the n-th value in the computation, because it is very close statistically in both worlds (that consider and "ignore" the n-th value).

Theorem 3. *(ϵ-differential privacy implies semantic privacy) Let $\mathcal{A}$ be an ϵ-differentially private algorithm. For all $D \in \mathcal{D}^n$ and $y \in \mathcal{Y}$, we have*

$$SD(b(D|y), b'(D|y)) \leq \exp(2\epsilon) - 1$$

Proof. By Bayes rule, we know that

$$b(D|y) = \frac{\mu(y|D)b(D)}{\sum\limits_{E \in \mathcal{D}^n} \mu(y|E)b(E)}$$

This yields

$$b(D|y) - b'(D|y) = \frac{\mu(y|D)b(D)}{\sum\limits_{E} \mu(y|E)b(E)} - \frac{\mu'(y|D)b(D)}{\sum\limits_{E} \mu'(y|E)b(E)}$$

Applying the inequalities of Definition 1, we get that $|b(D|y) - b'(D|y)| \leq \exp(2\epsilon) - 1$. □

5 Laplace mechanism

Now that we have discussed what differential privacy guarantees, the question remains, how do we realize differentially private algorithms? The most general mechanism is known as the Laplace mechanism [6].

First, let $f : \mathcal{D}^n \to \mathbb{R}^k$, and let $\|\cdot\|_1$ be the usual L_1 norm. Define $GS(f)$, the *global sensitivity* of f, for all neighboring databases D_1, D_2 as

$$GS(f) = \sup_{D_1,D_2 \in \mathcal{D}^n} \|f(D_1) - f(D_2)\|_1$$

Theorem 4. *(Laplace Mechanism [6]) Let f be defined as before and $\epsilon > 0$. Define randomized algorithm $\mathcal{A}$ as*

$$\mathcal{A}(D) = f(D) + Lap\left(\frac{GS(f)}{\epsilon}\right)^k$$

where the one-dimensional (zero mean) Laplace distribution $\mathrm{Lap}(b)$ *has density* $p(x; b) = \frac{1}{2b}\exp(-\frac{|x|}{b})$*, and $Lap(b)^k = (l_1, ..., l_k)$ where each $l_i \xleftarrow{iid} Lap(b)$. Then A is ϵ-differentially private.*

Proof. Let $y \in \mathbb{R}^k$ and D_1, D_2 be neighboring databases. Let $f(D)_i$ denote the i-th coordinate of $f(D)$. Then for each i we need to bound the ratio $\frac{p(y_i - f(D_1)_i)}{p(y_i - f(D_2)_i)}$.

$$\begin{aligned}
\frac{p(y_i - f(D_1)_i)}{p(y_i - f(D_2)_i)} &= \frac{\exp\left(-\epsilon \left|y_i - f(D_1)_i\right| / GS(f)\right)}{\exp\left(-\epsilon \left|y_i - f(D_2)_i\right| / GS(f)\right)} \\
&= \exp\left[\frac{\epsilon\left(|y_i - f(D_2)_i| - |y_i - f(D_1)_i|\right)}{GS(f)}\right] \\
&\le \exp\left[\frac{\epsilon\,|f(D_2)_i - f(D_1)_i|}{GS(f)}\right] \\
&\le \exp(\epsilon)
\end{aligned}$$

where the first inequality is triangle inequality and the second inequality comes from the definition of $GS(f)$. □

Note that if $z \sim Lap(b)$, then $\mathbb{E}[z] = 0$ and $\mathrm{Var}[z] = 2b^2$. So the Laplace mechanism adds noise with variance $2GS(f)^2/\epsilon^2$ along each coordinate. This agrees with our intuition, that (a) functions which have higher sensitivity require more noise to obtain a fixed ϵ of privacy, and (b) as we increase ϵ we can get away with adding less noise.

Noisy summation. Suppose $f(D) = \sum_{i=1}^{n} d_i$, where each $d_i \in \{0, 1\}$. Then clearly $GS(f) = 1$, so $\mathcal{A}(D) = \sum_{i=1}^{n} d_i + Lap(1/\epsilon)$ is an ϵ-differentially private version of sum. A natural question is, how much error does this approximate answer introduce? Since the Laplace distribution has zero mean, $\mathbb{E}[\mathcal{A}(D)] = f(D)$, so we can apply Chebyshev's inequality to get

$$\Pr\left[|\mathcal{A}(D) - f(D)| \ge k\right] \le \frac{2}{\epsilon^2 k^2}$$

So if $\epsilon = 0.1$, then we can be 95% sure that $\mathcal{A}(D)$ does not deviate from the actual answer by roughly 14.5. For large values of n, this seems quite reasonable.

Noisy average. Suppose $f(D) = \frac{1}{n}\sum\limits_{i=1}^{n} d_i$, where each $d_i \in [0, M]$ for some constant $M \in \mathbb{R}$. Then $GS(f) = M/n$, so $\mathcal{A}(D) = \frac{1}{n}\sum\limits_{i=1}^{n} d_i + Lap(\frac{M}{n\epsilon})$. Note that $GS(f) = O(1/n)$, which translates into another intuitive notion that adding more data yields more privacy.

Noisy linear regression. Suppose we have a dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, where each $x_i \in \mathbb{R}^k$ and $y_i \in \mathbb{R}$. Consider the least squares minimization function $f : D \to \mathcal{H}$

$$f(D) = \operatorname*{argmin}_{\theta \in \mathcal{H}} \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle \theta, x_i \rangle)^2 = \operatorname*{argmin}_{\theta \in \mathcal{H}} \hat{\mathcal{L}}(\theta; D)$$

where $\langle \cdot, \cdot \rangle$ is the standard inner product. Computing the exact global sensitivity of f is tricky, but we can derive an upper bound for it. Let $\|\cdot\|_2$ denote the standard L_2 norm. To make the analysis easier, we restrict ourselves to the case where $\|x_i\|_2 \leq M$, $|y_i| \leq N$, and $\mathcal{H} = \{\theta \in \mathbb{R}^k : \|\theta\|_2 \leq P\}$. Because f is a strongly convex function and $\mathcal{H}$ is a closed, convex set, we can apply the following fact from Chaudhuri et al. [3].

Theorem 5. *(Chaudhuri et al. [3]) If G and g are two strongly convex functions which are differentiable at all points, and if $\theta_1 = \operatorname{argmin}_\theta G(\theta)$ and $\theta_2 = \operatorname{argmin}_\theta G(\theta) + g(\theta)$, then*

$$\|\theta_1 - \theta_2\|_2 \leq \|\nabla g(\theta_2)\|_2$$

Let D_1, D_2 be two neighboring databases. Note that if we set $G(\theta) = \hat{\mathcal{L}}(\theta; D_1)$ and $g(\theta) = \hat{\mathcal{L}}(\theta; D_2) - \hat{\mathcal{L}}(\theta; D_1)$, then $\theta_1 = f(D_1)$ and $\theta_2 = f(D_2)$. We can now argue the following.

Theorem 6. *Let f be the least squares minimizer as above with all the given assumptions. Then $GS(f) \leq 2(Nk + PM^2\sqrt{k})/n$.*

Proof. We bound $\|\nabla g(\theta)\|_2$. Suppose wlog that D_1, D_2 differ in only the n-th element, and let $(x_n, y_n) \in D_1$ and $(x'_n, y'_n) \in D_2$. Then we have

$$\begin{aligned}\nabla g(\theta) &= \frac{1}{n}\left[(y_n - \langle \theta, x_n \rangle)x_n - (y'_n - \langle \theta, x'_n \rangle)x'_n\right] \\ &= \frac{1}{n}\left[y_n - y'_n + \langle \theta, x'_n \rangle x'_n - \langle \theta, x_n \rangle x_n\right]\end{aligned}$$

Taking the norm, we get

$$\begin{aligned}\|\nabla g(\theta)\|_2 &= \left\|\frac{1}{n}\left[y_n - y'_n + \langle \theta, x'_n \rangle x'_n - \langle \theta, x_n \rangle x_n\right]\right\|_2 \\ &\leq \frac{2N\sqrt{k}}{n} + \frac{1}{n}\left\|\langle \theta, x'_n \rangle x'_n\right\|_2 + \frac{1}{n}\left\|\langle \theta, x_n \rangle x_n\right\|_2 \\ &\leq \frac{2N\sqrt{k}}{n} + \frac{2PM^2}{n}\end{aligned}$$

Noting that $\|\cdot\|_1 \leq \sqrt{k}\,\|\cdot\|_2$ and applying Theorem 5 yields the result. □

Note: This mechanism is overly pessimistic in the amount of noise it adds to achieve ϵ-differential privacy. See [3] for more sophisticated techniques to achieve the same level of privacy while providing better utility. The basic idea is to modify the *objective* function instead of adding noise at the end. [3] also outlines a better ways to add noise for vector-valued functions which uses global sensitivity with respect to the L_2 norm instead of L_1.

6 Composition

One nice property of differential privacy that makes it much more practical is composibility.

Sequential composibility. The idea behind sequential composibility is that if we have k algorithms which are each independently differentially private, we would like to be able to feed the results from the first into the second, and so on, without completely sacrificing privacy. Sequential composibility allows us to do this.

More specifically, suppose we have k algorithms $\mathcal{A}_i(D; z_i)$, where the z_i represents some auxiliary input. Furthermore, suppose that each of the $\mathcal{A}_i$'s are ϵ-differentially private for any auxiliary input z_i. Consider a sequence of computations $\{z_1 = \mathcal{A}_1(D), z_2 = \mathcal{A}_2(D; z_1), z_3 = \mathcal{A}_3(D; z_1, z_2), ...\}$, and suppose $\mathcal{A}(D) = z_k$.

Theorem 7. *(Sequential composibility [13]) $\mathcal{A}(D)$ is $k\epsilon$-differentially private.*

Proof. Let D_1, D_2 be two neighboring databases. Then

$$\begin{aligned}\mu(z_k|D_1) &= \mu(z_1|D_1)\mu(z_2|D_1, z_1)...\mu(z_k|D_1, z_1, ..., z_{k-1}) \\ &\leq \exp(k\epsilon)\prod_{i=1}^{k}\mu(z_i|D_2, z_1, ..., z_{i-1}) \\ &= \exp(k\epsilon)\mu(z_k|D_2)\end{aligned}$$

□

Sequential composibility is very useful for iterative algorithms which run over the same dataset multiple times. If we can make each iteration differentially private, and we can bound the number of iterations needed, then we can appeal to sequential composibility to argue the entire process is differentially private.

Parallel composibility. Now consider the situation where we have a single database D partitioned into k disjoint subsets, D_i. Once again, suppose we have k algorithms $\mathcal{A}_i(D_i; z_i)$ which are each ϵ differentially private. Once again, suppose $\mathcal{A}(D) = z_k$.

Theorem 8. *(Parallel composibility [13]) $\mathcal{A}(D)$ is ϵ-differentially private.*

Proof. Let D_1, D_2 be two neighboring databases. Suppose that the j-th partition contains the differing element. Then

$$\begin{aligned}\mu(z_k|D_1) &= \prod_{i=1}^{k} \mu(z_i|D_{1_i}, z_1, ..., z_{i-1}) \\ &\leq \exp(\epsilon)\mu(z_j|D_{2_j}, z_1, ..., z_j) \prod_{i \neq j}^{k} \mu(z_i|D_{1_i}, z_1, ..., z_{i-1}) \\ &= \exp(\epsilon)\mu(z_k|D_2)\end{aligned}$$

□

7 Beyond the basics

We have barely begun to scratch the surface of all the work done in the area of differential privacy. Below are the broad areas of research people are doing in differential privacy, with a few (not at all comprehensive) sample papers

Relaxed definitions. ϵ-differential privacy is a very strong, worst case definition. A lot of researchers have considered various relaxations to the definition to allow for algorithms to achieve better utility. See e.g. [10, 1, 11].

Applying differential privacy to algorithms. A lot of clever techniques have been used to produce algorithms which achieve differential privacy without having to add worst case noise. Various machine learning algorithms, such as decision trees, SVMs, logistic regression have differentially private variants which are practical. See e.g. [3, 12, 7, 9].

Theoretical noise requirements and utility bounds. A lot of theoretical analysis has also been done to answer questions such as how much noise must be added to classes of algorithms to achieve differential privacy, and how much utility must differentially private algorithms give up? See e.g. [8, 4, 2].

References

[1] R. Bassily, A. Groce, J. Katz, and A. Smith. Coupled-worlds privacy: exploiting adversial uncertainty in statistical data privacy. FOCS, 2013.

[2] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th annual ACM symposium on Theory of computing*, STOC '08, 2008.

[3] K. Chaudhuri, C. Monteleoni, and A. D. Sarwate. Differentially private empirical risk minimization. *J. Mach. Learn. Res.*, 12:1069–1109, 2011.

[4] A. De. Lower bounds in differential privacy. *CoRR*, abs/1107.2183, 2011.

[5] C. Dwork. Differential privacy. In *ICALP*, pages 1–12, 2006.

[6] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third conference on Theory of Cryptography*, TCC'06, pages 265–284, 2006.

[7] R. Hall, A. Rinaldo, and L. Wasserman. Differential privacy for functions and functional data. *J. Mach. Learn. Res.*, 14(1), 2013.

[8] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM symposium on Theory of computing*, STOC '10, pages 705–714, 2010.

[9] G. Jagannathan, K. Pillaipakkamnatt, and R. N. Wright. A practical differentially private random decision tree classifier. In *Proceedings of the 2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, 2009.

[10] S. P. Kasiviswanathan and A. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, abs/0803.3946, 2008.

[11] D. Kifer and A. Machanavajjhala. A rigorous and customizable framework for privacy. In *Proceedings of the 31st symposium on Principles of Database Systems*, PODS '12, pages 77–88, 2012.

[12] D. Kifer, A. D. Smith, and A. Thakurta. Private convex optimization for empirical risk minimization with applications to high-dimensional regression. *Journal of Machine Learning Research - Proceedings Track*, 23, 2012.

[13] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data*, SIGMOD '09, pages 19–30, 2009.

[14] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.