Information-theoretic foundations of differential privacy*

Darakhshan J. Mir

Rutgers University, Piscataway NJ 08854, USA,
mir@cs.rutgers.edu

Abstract. We examine the information-theoretic foundations of the increasingly popular notion of *differential privacy*. We establish a connection between differential private mechanisms and the *rate-distortion* framework. Additionally, we also show how differentially private distributions arise out of the application of the *Maximum Entropy Principle*. This helps us locate differential privacy within the wider framework of information-theory and helps formalize some intuitive aspects of our understanding of differential privacy.

1 Introduction

The problem of releasing aggregate information about a statistical database while simultaneously providing privacy to the individual participants of the database has been extensively studied in the computer science and statistical communities. *Differential privacy* (DP) has been one of the main lines of research that has emerged out of attempts to formalize and solve this problem, over the last few years. See [5] for a survey. It formalizes the idea that privacy is provided if the "identification risk" an individual faces does not change appreciably if he or she participates in a statistical database.

Often, in the context of data privacy, and more specifically, differential privacy, the claim is made that *privacy* and *utility* are conflicting goals. The application of differential privacy to several problems of private data analysis has made it clear that the utility of the data for a specific measurement degrades with the level of privacy. The greater the level of privacy, the less "useful" the data is, and vice versa. This paper attempts to understand the precise information-theoretic conditions that necessitate such a trade-off. We observe that differentially-private mechanisms arise out of minimizing the information leakage (measured using information-theoretic notions such as mutual information) while trying to maximize "utility". The notion of utility is captured by the use of an abstract distortion function **dist** that measures the distortion between the input and the output of the mechanism. This is a general mechanism, and can be instantiated appropriately depending on the problem domain. The main observation of this paper is that the probability distribution that achieves this constrained minimization corresponds to the so-called *exponential mechanism* [11]. We also show

* This work was supported by NSF award number CCF-1018445.

how differentially-private mechanisms arise out of the application of the *principle of maximum entropy*, first formulated by Jaynes [7]. We see that among all probability distributions that constrain the expected distortion to stay within a given value, the differentially private mechanism, corresponds to the distribution that maximizes the conditional entropy of output given the input. This, to our knowledge, is the first attempt at providing an information theoretic foundation for differential privacy. In Section 2 we review the appropriate definitions and notions from differential privacy. In Section 2.1 we discuss related work. In Sections 3 and 4 we present our main results.

2 Definitions and Background

In this section we present the background and the related work in differential privacy. Assume a probability distribution $p_{\mathbf{X}}(\mathbf{x})$ on an alphabet $\mathcal{X}$. $\mathcal{X}$ may either be a scalar or vector space. Let $\mathbf{X}_i \in \mathcal{X}$ be a random variable representing the i-th row of a database. Then the random variable representing a database of size n, (whose elements are drawn from $\mathcal{X}$) is $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2 \ldots, \mathbf{X}_n)$. $\mathbf{x}$ represents the value that the random variable $\mathbf{X}$ takes, that is the observed database $\mathbf{x}$. Note that the $\mathbf{X}$'s themselves may multi-dimensional representing the k attributes of the database. Dwork et al. [6] define the notion of differential privacy that provides a guarantee that the probability distribution on the outputs of a mechanism is "almost the same," irrespective of whether or not an individual is present in the data set. Such a guarantee incentivizes participation of individuals in a database by assuring them of incurring very little risk by such a participation. To capture the notion of a user opting in or out, the "sameness" condition is defined to hold with respect to a neighbor relation; intuitively, two inputs are neighbors if they differ only in the participation of a single individual. For example, Dwork et al. [6] define datasets to be neighbors if they differ in a single row. McGregor et. al [10] define differential privacy, equivalently, in terms of probability distributions. This formulation is more useful for us.

Definition 1. *[10] Let* $\mathbf{x}$ *be a database of length* n*, drawing each of its elements from an alphabet* $\mathcal{X}$*, then an* ε-differentially private mechanism *on* $\mathcal{X}^n$ *is a family of probability distributions* $\{\pi(\mathbf{o}|\mathbf{x}) : \mathbf{x} \in \mathcal{X}^n\}$ *on a range* $\mathcal{O}$*, such that for every neighboring* $\mathbf{x}$ *and* $\mathbf{x}'$*, and for every measurable subset* $\mathbf{o} \subset \mathcal{O}$*,* $\pi(\mathbf{o}|\mathbf{x}) \leq \pi(\mathbf{o}|\mathbf{x}') \exp(\varepsilon)$.

Notice that the distribution (or equivalently) mechanism is parametrized by the input database $\mathbf{x}$ or $\mathbf{x}'$, whichever is relevant.

One mechanism that Dwork et al. [6] use to provide differential privacy is the *Laplacian noise method* which depends on the *global sensitivity* of a function:

Definition 2. *[6] For* $f : \mathcal{X}^n \to \mathbb{R}^d$*, the global sensitivity of* f *is* $\Delta f = \max_{\mathbf{x} \sim \mathbf{x}'} ||f(\mathbf{x}) - f(\mathbf{x}')||_1$.

Another, more general (though, not always computationally efficient) method of providing differential privacy is the so called *exponential mechanism* proposed

by McSherry and Talwar [11]. This mechanism can be said to be parametrized by a "distortion function" $\mathbf{dist}(\mathbf{x}, \mathbf{o})$ that maps a pair of an input data set $\mathbf{x}$ (a vector over some arbitrary real-valued domain) and candidate output $\mathbf{o}$ (again over an arbitrary range $\mathcal{O}$) to a real valued "distortion score." Lower valued distortions imply good input-output correspondences. It assumes a base measure π on the range $\mathcal{O}$. For a given input $\mathbf{x}$, the mechanism selects an output $\mathbf{o}$ with exponential bias in favor of low distorting outputs by sampling from the following *exponential distribution* [11]:

$$\pi^{\varepsilon}(\mathbf{o}) \propto \exp(-\varepsilon\, \mathbf{dist}(\mathbf{x}, \mathbf{o})) \cdot \pi(\mathbf{o}). \tag{1}$$

Theorem 1. *[11] The exponential mechanism, when used to select an output* $\mathbf{o} \in \mathcal{O}$*, gives* $2\varepsilon\Delta\,\mathbf{dist}$*-differential privacy, where* $\Delta\,\mathbf{dist}$ *is the global sensitivity of the distortion function* $\mathbf{dist}$*.*

The exponential mechanism is a useful abstraction when trying to understand differential privacy because it generalizes all specific mechanisms, such as the Laplacian mechanism introduced above. The exponential mechanism because of the generality of the input space $\mathcal{X}$, the output range $\mathcal{O}$ and the distortion function $\mathbf{dist}$, captures all differentially private mechanisms. The π^{ε} denotes the dependence of the posterior on $\pi(\mathbf{o}|\mathbf{x})$, on the parameter ε.

2.1 Related work

Some information-theoretic notions and metrics of data privacy exist in the literature. See [17], [3], for example. Sankar et. al [14] consider the problem of quantifying the privacy risk and utility of a data transformation in an information-theoretic framework. Rebello-Monedero [13] consider the problem in a similar framework and define an information-theoretic privacy measure similar to an earlier defined measure of *t-closeness* [8]. A connection between information theory and differential privacy through Quantitative flow has been made by Alvim et al. [1]. Alvim et al. [1] use the information-theoretic notion of Min-entropy for the information leakage of the private channel, and show that differential privacy implies a bound on the min-entropy of such a channel. They also show how differential privacy imposes a bound on the utility of a randomized mechanism and under certain conditions propose an optimal randomization mechanism that achieves a certain level of differential privacy. Barthe and Kopf [2] also develop upper bounds for the leakage of every ε-differentially private mechanism. Our work is different from (but related to) theirs in the sense that we do not aim at finding bounds for the information leakage (or risk) of the differentially-private mechanisms. Our aim is to understand the information-theoretic foundations of the framework of differential privacy. Our work is in the spirit of Sankar et al. [14] and Rebello-Monedero et al. [13] but examining how a risk-distortion tradeoff gives rise to differentially-private mechanisms. In previous work [12] we examine the information theoretic connections of differentially-private learning. This was done in a specific context of learning, and the general implications were not clear.

3 Differentially-private mechanisms in a risk-distortion framework

Assume an input space $\mathcal{X}^n$, and a range $\mathcal{O}$. For any $\mathbf{x} \in \mathcal{X}^n$, and any output $\mathbf{o} \in \mathcal{O}$, a distortion function **dist** is specified. Consider a probability measure $p_{\mathbf{X}}(\mathbf{x})$ on $\mathcal{X}$ and a prior probability π on $\mathcal{O}$.

Given a database $\mathbf{x}$, which is a set of n random independent samples $\mathbf{X} = \{\mathbf{X}_1, \dots \mathbf{X}_n\} \in \mathcal{X}^n$, where each $\mathcal{X}_i$ is drawn i.i.d from $p_{\mathbf{X}}(\mathbf{x})$, and an output $\mathbf{o}$, the "utility" of $\mathbf{o}$ for $\mathbf{x}$, is given by (the negative of) a function $\mathbf{dist} : \mathcal{X}^n \times \mathcal{O} \to \mathbb{R}$.

The expected distortion of a mechanism $\pi_{\mathbf{O}|\mathbf{X}}(\mathbf{o}|\mathbf{x})$ is:

$\mathbb{E}_{\mathbf{x} \sim p_{\mathbf{X}}(\mathbf{x})^n} \mathbb{E}_{\mathbf{o} \sim \pi(\mathbf{o}|\mathbf{x})}\, \mathbf{dist}(\mathbf{x}, \mathbf{o})$.

Rebollo-Monedero et. al [13] define a privacy risk function to be the mutual information between the revealed and the hidden random variables. Similarly, we define a privacy risk function $\mathcal{R}$ to be the mutual information between the input (the underlying database) and the output of the differentially private mechanism, that is, $\mathcal{R} = I(\mathbf{X}; \mathbf{O})$. We know that the mutual information

$$I(\mathbf{X}; \mathbf{O}) = H(\mathbf{O}) - H(\mathbf{O}|\mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X}|\mathbf{O}), \tag{2}$$

where $H(\mathbf{X})$ represents the entropy of the random variable of $\mathbf{X}$ and $H(\mathbf{O}|\mathbf{X})$ the conditional entropy of $\mathbf{O}$ given $\mathbf{X}$. So, the mutual information is the reduction in the uncertainty about $\mathbf{X}$ by knowledge of output $\mathbf{O}$ or vice versa (See [4] for example). Also we have that

$$\mathcal{R} = I(\mathbf{X}; \mathbf{O}) = \mathbb{E} \log \frac{\pi(\mathbf{O}|\mathbf{X}) p(\mathbf{X})}{\pi(\mathbf{O}) p(\mathbf{X})} = \mathbb{E} \log \frac{\pi(\mathbf{O}|\mathbf{X})}{\pi(\mathbf{O})}. \tag{3}$$

This is equal to the conditional Kullback-Leibler divergence between the posterior and prior distributions denoted by $D_{KL}\left(\pi(\mathbf{O}|\mathbf{X}) \| \pi(\mathbf{O})\right)$. If the prior and posterior distributions are the same, then the privacy risk is zero, but that also means that the distortion may be arbitrarily high. However, we are interested in minimizing the distortion function associated with the posterior distribution, while minimizing the privacy risk $\mathcal{R}$. As a result, we are interested in quantifying this risk-distortion trade-off. Notice that untill this point, our risk-distortion framework is formulated only in information-theoretic terms. We will see how the differentially-private mechanism arises out of this framework.

As in Rebollo-Mondero et. al [13], we are interested in a randomized output, minimizing the privacy risk given a distortion constraint (or viceversa). Unlike their treatment, however, the potential outputs are more general than perturbations of the input database elements to capture differentially-private mechanims (both interactive and noninteractive). The privacy risk-distortion function is defined analogously (as in Rebollo-Mondero [13]), as

$$\mathcal{R}(\mathcal{D}) = \inf_{\pi_{\mathbf{O}|\mathbf{X}} \,:\, \mathbb{E}_{\mathbf{x},\mathbf{o}}\, \mathbf{dist}(\mathbf{x}, \mathbf{o}) \leq \mathcal{D}} I(\mathbf{X}; \mathbf{O}) \tag{4}$$

3.1 Connection to the rate-distortion framework

Rebollo-Mondero et. al relate the risk-distortion function formulated in Equation 4 [13] to the well-known *rate-distortion* problem in information theory first formulated by Shannon. (See [4], for example). Shannon's rate-distortion theory is applied in the context of lossy compression of data. The objective is to construct a compact representation (a code) of the underlying signal (or data), such that the average distortion of the signal reconstructed from this compact representation is low. Rate-distortion theory determines the level of expected distortion $\mathcal{D}$, given the desired information rate $\mathcal{R}$ of the code or vice-versa using the rate-distortion function $\mathcal{R}(\mathcal{D})$ similar to that in Equation 4 where $\mathcal{R}$ is the information rate of the code, when applied to the compression problem. So, the rate-distortion function is defined as the infimum of the rates of codes whose distortion is bounded by $\mathcal{D}$.

Using this connection, one can prove the following:

Theorem 2. *[13] The privacy risk-distortion function is a convex and non-increasing function of $\mathcal{D}$.*

The problem is to minimize the privacy risk, defined thus, under the expected distortion constraint. As a function of the probability density, $\pi_{\mathbf{O}|\mathbf{X}}(\mathbf{o}|\mathbf{x})$, the problem is also convex. We can also use Lagrangian multipliers to write Equation 4 in an equivalent unconstrained form. We have the functional

$$\mathcal{F}[\pi(\mathbf{o}|\mathbf{x})] = \frac{1}{\varepsilon} I(\mathbf{X};\mathbf{O}) + \mathbb{E}\,\mathbf{dist}(\mathbf{X},\mathbf{O}). \tag{5}$$

for a positive ε. Functional $\mathcal{F}$ needs to be minmized among all normalized $\pi(\mathbf{o}|\mathbf{x})$. So we can find the distribution that minimizes this funtion, by using standard optimization techniques. Standard arithmetic manipulation, leads Tishby et al. [16] to prove the following theorem:

Theorem 3. *[16] The solution of the variational problem, $\frac{\partial \mathcal{F}}{\partial \pi(\mathbf{o}|\mathbf{x})} = 0$, for normalized distributions $\pi(\mathbf{o}|\mathbf{x})$, is given by the exponential form*

$$\pi^{\varepsilon}(\mathbf{o}|\mathbf{x}) = \frac{\exp(-\varepsilon\,\mathbf{dist}(\mathbf{x},\mathbf{o}))}{Z(\mathbf{x})}\pi(\mathbf{o}). \tag{6}$$

where $Z(\mathbf{x},\varepsilon)$ is a normalization (partition) function. Moreover, the Lagrange multiplier ε is determined by the value of the expected distortion, $\mathcal{D}$, is positive and satisfies, $\frac{\partial \mathcal{R}}{\partial \mathcal{D}} = -\varepsilon$.

We have that among all the conditional distributions, the one that optimizes this functional in Equation 5 is π^{ε} in Equation 6 above. This is our main result, that the distribution that minimizes the privacy risk, given a distortion constraint is a differentially-private distribution. From examining equation 1 and Theorem 1 we have

Theorem 4. *The distribution that minimizes Equation 4 defines a $2\varepsilon\Delta\,\mathbf{dist}$-differentially private mechanism.*

Figure 1 illustrates the tradeoff. It plots the unconstrained Lagrangian function $L(\mathcal{D}, \mathcal{R}) = \mathcal{D} + \frac{1}{\varepsilon}\mathcal{R}$, which because of the convexity of the risk-distortion function is also convex. For a given privacy parameter ε, we consider lines of slope $-\varepsilon$. We see that these lines intersect the curve at various points, these points represent the risk-distortion tradeoffs for those values. As we should expect, a high privacy-risk implies a low distortion and vice-versa. We see that for a given value of $-\varepsilon$, the line that is tangent to the curve at represents the optimal tradeoff point between the risk and the distortion. The value of the function $L(\mathcal{D}, \mathcal{R})$ on these lines is a constant, which implies that in some way the level of privacy imposes a value on the function L, since such a line can only intersect the curve in at most two places.

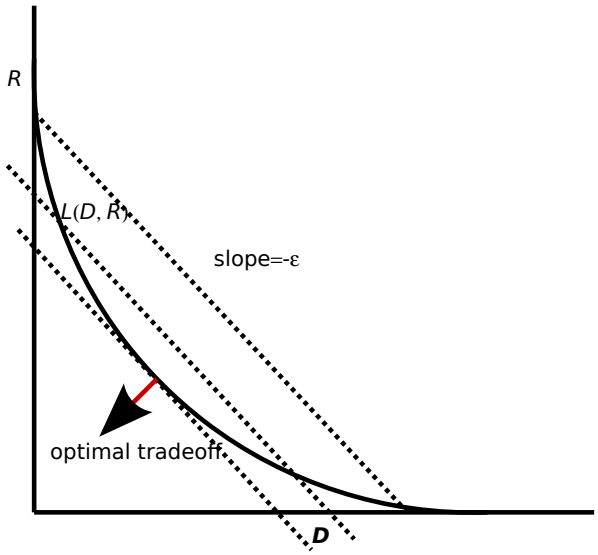

Fig. 1: Risk-distortion curve

4 Differential privacy arising out of the Maximum Entropy principle or Minimum Discrimination Information principle

The *principle of maximum entropy* was proposed by Jaynes [7]. Suppose, a random variable $\mathbf{X}$ takes a discrete set of values $\mathbf{x}_i$ with probabilities specified by $p_{\mathbf{X}}(\mathbf{x}_i)$, and we know of constraints on the distribution $p_{\mathbf{X}}$, in the form of expectations of some functions of these random variables. Then the principle of maximum entropy states that of all distributions $p_{\mathbf{X}}$ that satisfy the constraints, one should choose the one with the largest entropy $H(\mathbf{X}) = -\sum_i p(\mathbf{x}_i) \log(p(\mathbf{x}_i))$.

In the case of a continuous random variable, the Shannon entropy is not useful and for such cases we apply the principle of minimum discrimination information [7]. It states that given a prior p on $\mathbf{X}$, a new distribution q should be chosen so that it as hard as possible to distinguish it from the prior distribution p, that is the new data should produce as small a gain in information as possible given by $D_{KL}(q\|p)$.

We show that the application of the principle of Maximum Entropy to the distribution $\pi(\mathbf{o}|\mathbf{x})$ gives rise to a differentially-private mechanism.

When trying to find a distribution $\pi_{\mathbf{O}|\mathbf{X}}(\mathbf{o}|\mathbf{x})$, we utilize the Maximum Entropy Principle. Among all distributions $p(\mathbf{o}|\mathbf{x})$, we choose the one that maximizes the entropy $H(\mathbf{O}|\mathbf{X})$ subject to satisfying the constraint that the expected distortion function $\mathbf{dist}(\mathbf{o},\mathbf{x})$ is bounded by a quantity D. So we have,

$$\textbf{maximize } H(\mathbf{O}|\mathbf{X})$$
$$\textbf{subject to } \sum \mathbf{dist}(\mathbf{x},\mathbf{o})p(\mathbf{o}|\mathbf{x})p(\mathbf{x}) \leq \mathcal{D}.$$

From equation 2 we observe that minimizing the mutual information as in Equation 4 is equivalent to maximizing the entropy $H(\mathbf{O}|\mathbf{X})$.

Shannon introduced the concept of *equivocation* as the conditional entropy of a private message given the observable [15]. Sankar et. al [14] use equivocation as a measure of privacy of their data transformation. Their aim is also to maximize the average equivocation of the underlying secret sample given the observables Since $I(\mathbf{X};\mathbf{O}) = H(\mathbf{X}|\mathbf{O}) - H(\mathbf{X})$, minimizing $I(\mathbf{X};\mathbf{O})$ is also equivalent to maximizing the conditional entropy $H(\mathbf{X}|\mathbf{O})$, subject to constraints on the expected distortion. Therefore, the exponential distribution $\pi^{\varepsilon}(\mathbf{o}|\mathbf{x})$ as defined in Equation 6 maximizes the conditional uncertainty about the underlying sample given a constraint on the distortion function.

Now consider the worst case which differential privacy protects against, that is given knowledge of the entire database except for one row i, represented as $\mathbf{X}_{-i}$, if we look at the problem of maximizing the uncertainty of the random variable $\mathbf{X}_i$, we have

$$\text{maximize } H(\mathbf{X}_i|\mathbf{O},\mathbf{X}_{-i})$$
$$\text{subject to } \sum \mathbf{dist}(x_i,\mathbf{x}_{-i},\mathbf{o})p(x_i|\mathbf{x}_{-i},\mathbf{o})p(\mathbf{x}_{-i},\mathbf{o}) \leq \mathcal{D}$$

Again this is equivalent to minimizing the mutual information $I(\mathbf{X},\mathbf{O})$ when $\mathbf{X}_{-i}$ and $\mathbf{O}$ are given.

A note on incorporating auxilliary information: Usually, differential privacy provides guarantees on the inference, irrespective of any side or auxilliary information. This can be easily incorporated in our framework like Sankar et. al [14] by making all the distributions above conditional on the side information.

5 Conclusion and future work

We presented an information-theoretic foundation for differential privacy, which to our knowledge is the first such attempt. We formulated differential privacy

within the broader frameworks of various problems in information theory such as the rate-distortion problem and the maximum entropy principle. There are several directions for future work.

One, we can try to apply the risk-distortion framework to examine the generation of private synthetic data when the underlying data generating distribution $p_{\mathbf{X}}(\mathbf{x})$ is known. Additionally, one could try derive bounds on the mutual information in such cases. Second, we can examine the deployment of this framework to problems where the distortion function **dist** is specified. Another direction is to examine the notion of compressive privacy [9] in this rate-distortion framework and derive bounds for the rate.

References

1. M. S. Alvim, M. E. Andrés, K. Chatzikokolakis, P. Degano, and C. Palamidessi. Differential privacy: On the trade-off between utility and information leakage. In *FST '11*.
2. G. Barthe and B. Kopf. Information-theoretic bounds for differentially private mechanisms. In *CSF '11*.
3. M. Bezzi. An information theoretic approach for privacy metrics. *TDP '10*, 3(3):199–215.
4. T. M. Cover and J. A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.
5. C. Dwork. Differential privacy. In *ICALP(2) '06*.
6. C. Dwork, F. Mcsherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC '06*.
7. E. T. Jaynes. Information theory and statistical mechanics. ii. *Phys. Rev. '57*, 108:171–190.
8. N. Li and T. Li. t-closeness: Privacy beyond k-anonymity and -diversity. In *ICDE 07*.
9. Y. D. Li, Z. Zhang, M. Winslett, and Y. Yang. Compressive mechanism: Utilizing sparse representation in differential privacy. *CoRR*, abs/1107.3350, 2011.
10. A. McGregor, I. Mironov, T. Pitassi, O. Reingold, K. Talwar, and S. P. Vadhan. The limits of two-party differential privacy. In *FOCS '10*.
11. F. Mcsherry and K. Talwar. Mechanism design via differential privacy. In *FOCS '07*.
12. D. Mir. Differentially-private learning and information theory. In *, EDBT-ICDT-W '12*.
13. D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer. From t-closeness-like privacy to postrandomization via information theory. *IEEE TKDE '10*, 22:1623–1636.
14. L. Sankar, S. Rajagopalan, and H. Poor. A theory of utility and privacy of data sources. In *ISIT '10*.
15. C. Shannon. Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, Part 4*, pages 142–163, 1959.
16. N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. *Allerton '99*.
17. P. L. Vora. An information-theoretic approach to inference attacks on random data perturbation and a related privacy measure. *IEEE Trans. Inf. Theor. '07*, 53(8):2971–2977.