Bayesian Differential Privacy for Machine Learning

Aleksei Triastcyn [1] **Boi Faltings** [1]

Abstract

Traditional differential privacy is independent of the data distribution. However, this is not well-matched with the modern machine learning context, where models are trained on specific data. As a result, achieving meaningful privacy guarantees in ML often excessively reduces accuracy. We propose *Bayesian differential privacy (BDP)*, which takes into account the data distribution to provide more practical privacy guarantees. We also derive a general privacy accounting method under BDP, building upon the well-known moments accountant. Our experiments demonstrate that in-distribution samples in classic machine learning datasets, such as MNIST and CIFAR-10, enjoy significantly stronger privacy guarantees than postulated by DP, while models maintain high classification accuracy.

1. Introduction

Machine learning (ML) and data analytics offer vast opportunities for companies, governments and individuals to take advantage of the accumulated data. However, their ability to capture fine levels of detail can compromise privacy of data providers. Recent research (Fredrikson et al., 2015; Shokri et al., 2017; Hitaj et al., 2017) suggests that even in a black-box setting it is possible to infer information about individual records in the training set.

Numerous solutions have been proposed to address this problem, varying in the extent of data protection and how it is achieved. In this work, we consider a notion that is viewed by many researchers as the gold standard – *differential privacy (DP)* (Dwork, 2006). Initially, DP algorithms focused on sanitising simple statistics, but in recent years, it made its way to machine learning (Abadi et al., 2016; Papernot et al., 2016; 2018; McMahan et al., 2018).

[1]Artificial Intelligence Lab, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. Correspondence to: Aleksei Triastcyn <aleksei.triastcyn@epfl.ch>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

Despite notable advances, differentially private ML still suffers from two major problems: (a) utility loss due to excessive noise added during training and (b) difficulty in interpreting the privacy parameters ε and δ. In many cases where the first problem appears to be solved, it is actually being hidden by the second. We design a motivational example in Section 3 that illustrates how a seemingly strong privacy guarantee allows for the attacker accuracy to be as high as 99%. Although this guarantee is very pessimistic and holds against a powerful adversary with any auxiliary information, it can hardly be viewed as a reassurance to a user. Moreover, it provides only the worst-case bound, leaving users to wonder how far is it from a typical case.

In this paper, we focus on practicality of privacy guarantees and propose a variation of DP that provides more meaningful guarantees for *typical* scenarios on top of the global DP guarantee. We name it *Bayesian differential privacy (BDP)*.

The key to our privacy notion is the definition of *typical* scenarios. We observe that machine learning models are designed and tuned for a particular data distribution (for example, an MRI dataset is very unlikely to contain a picture of a car). Moreover, such prior distribution of data is often already available to the attacker. Thus, we consider a scenario *typical* when all sensitive data is drawn from the same distribution. While the traditional differential privacy treats all data as equally likely and hides differences by large amounts of noise, Bayesian DP calibrates noise to the data distribution and provides much tighter expected guarantees.

As the data distribution is usually *unknown*, BDP estimates the necessary statistics from data as shown in the following sections. Furthermore, since typical scenarios are determined by data, the participants of the dataset are covered by the BDP guarantee with high probability.

To accompany the notion of Bayesian DP (Section 4.1), we provide its theoretical analysis and the privacy accounting framework (Section 4.2). The latter considers the privacy loss random variable and employs principled tools from probability theory to find concentration bounds on it. It provides a clean derivation of privacy accounting in general (Sections 4.2 and 4.3), as well as in the special case of subsampled Gaussian noise mechanism. Moreover, we show that it is a generalisation of the well-known moments accountant (MA) (Abadi et al., 2016) (Section 4.4.2).

Since our privacy accounting relies on data distribution samples, a natural concern is that the data not present in the dataset are not taken into account, and thus, are not protected. However, our finite sample estimator is specifically designed to address this issue (see Section 4.3).

Our contributions in this paper are the following:

- we propose a variation of DP, called Bayesian differential privacy, that allows to provide more practical privacy guarantees in a wide range of scenarios;
- we derive a clean, principled privacy accounting method that generalises the moments accountant;
- we experimentally demonstrate advantages of our method (Section 5), including the state-of-the-art privacy bounds in deep learning (Section 5.2).

2. Related Work

With machine learning applications becoming more and more ubiquitous, vulnerabilities and attacks on ML models get discovered, raising the need for matching defences. These attacks can be based on both passive adversaries, such as model inversion (Fredrikson et al., 2015) and membership inference (Shokri et al., 2017), and active adversaries (for example, (Hitaj et al., 2017)).

One of the strongest privacy standards that can be employed to protect ML models from these and other attacks is differential privacy (Dwork, 2006; Dwork et al., 2006). Pure ε-DP is hard to achieve in many realistic learning settings, and therefore, a notion of approximate (ε, δ)-DP is used across-the-board in machine learning. It is typically accomplished by applying the Gaussian noise mechanism (Dwork et al., 2014) during the gradient descent update (Abadi et al., 2016). Privacy accounting, i.e. computing the privacy guarantee throughout multiple iterations of the algorithm, is typically done by the *moments accountant (MA)* (Abadi et al., 2016). In Section 4.4.2, we discuss the link between MA and our accounting method, as well as connection to a closely related notion of Rényi DP (Mironov, 2017). Similarly, a link can be established to concentrated DP definitions (Dwork & Rothblum, 2016; Bun & Steinke, 2016).

A number of previous relaxations considered a similar idea of limiting the scope of protected data or using the data generating distribution, either through imposing a set of data evolution scenarios (Kifer & Machanavajjhala, 2014), policies (He et al., 2014), distributions (Blum et al., 2013; Bhaskar et al., 2011), or families of distributions (Bassily et al., 2013; Bassily & Freund, 2016). Some of these definitions (e.g. (Blum et al., 2013)) may require more noise, because they are stronger than DP in the sense that datasets can differ in more than one data point. This is not the case with our definition: like DP, it considers adjacent datasets *differing in a single data point*. The major problem of such definitions, however, is that in real-world scenarios it is not feasible to exactly define distributions or families of distributions that generate data. And even if this problem is solved by restricting the query functions to enable the usage of the central limit theorem (e.g. (Bhaskar et al., 2011; Duan, 2009)), these guarantees would only hold asymptotically and may require prohibitively large batch sizes. While Bayesian DP can be seen as a special case of some of the above definitions, the crucial difference with the prior work is that our additional assumptions allow the Bayesian accounting (Sections 4.2, 4.3) to provide guarantees w.h.p. with finite number of samples from data distributions, and hence, enable a broad range of real-world applications.

Finally, there are other approaches that use the data distribution information in one way or another, and coincidentally share the same (Yang et al., 2015) or similar (Leung & Lui, 2012) names. Yet, similarly to the methods discussed above, their assumptions (e.g. bounds on the minimum probability of a data point) and implementation requirements (e.g. potentially constructing correlation matrices for millions of data samples) make practical applications very difficult. Perhaps, the most similar to our approach is random differential privacy (Hall et al., 2011). The main difference is that Hall et al. (2011) consider the probability space over all data points, while we only consider the space over a single differing example. As a result, our guarantees are more practical to compute for large, realistic ML datasets. Furthermore, Hall et al. (2011) only propose a basic composition theorem, which is not tight enough for accounting in iterative methods, and to the best of our knowledge, there are no proofs for other crucial properties, such as post-processing and group privacy.

3. Motivation

Before we proceed, we find it important to motivate research on alternative privacy definitions, as opposed to fully concentrating on new mechanisms for DP. On the one hand, there is always a combination of data and a desired statistic that would yield large privacy loss in DP paradigm, regardless of the mechanism. In other words, there can always be data outliers that are difficult to hide without a large drop in accuracy. On the other hand, we cannot realistically expect companies to sacrifice model quality in favour of privacy. As a result, we get models with impractical worst-case guarantees (as we demonstrate below) without any indication of what is the privacy guarantee for the majority of users.

Consider the following example. The datasets D, D' consist of income values for residents of a small town. There is one individual x' whose income is orders of magnitude higher than the rest, and whose residency in the town is what the attacker wishes to infer. The attacker observes the mean income w sanitised by a differentially private

mechanism with $\varepsilon = \varepsilon_0$ (we consider the stronger, pure DP for simplicity). What we are interested in is the change in the posterior distribution of the attacker after they see the private model compared to prior (Mironov, 2017; Bun, 2017). If the individual is not present in the dataset, the probability of w being above a certain threshold is extremely small. On the contrary, if x' is present, this probability is higher (say it is equal to r). The attacker computes the likelihood of the observed value under each of the two assumptions, the corresponding posteriors given a flat prior, and applies a Bayes optimal classifier. The attacker then concludes that the individual is present in the dataset and is a resident.

By the above expression, r can only be e^{ε_0} times larger than the respective probability without x'. However, if the $re^{-\varepsilon_0}$ is small enough, then the probability $P(A)$ of the attacker's guess being correct is as high as $r/(r + re^{-\varepsilon_0})$, or

$$P(A) = \frac{1}{1 + e^{-\varepsilon}}. \tag{1}$$

To put it in perspective, for a DP algorithm with $\varepsilon = 2$, the upper bound on the accuracy of this attack is as high as 88%. For $\varepsilon = 5$, it is 99.33%. For $\varepsilon = 10$, 99.995%. Importantly, these values of ε are very common in DP ML literature (Shokri & Shmatikov, 2015; Abadi et al., 2016; Papernot et al., 2018), and they can be even higher in real-world deployments[1].

This guarantee does not tell us anything other than that this outlier cannot be protected while preserving utility. But what is the guarantee for other residents of the town? Intuitively, it should be much stronger. In the next section, we present a novel DP-based privacy notion. It uses the same privacy mechanism and augments the general DP guarantee with a much tighter guarantee for the expected case, and, by extension, for any percentile of the user/data population.

4. Bayesian Differential Privacy

In this section, we define *Bayesian differential privacy (BDP)*. We then derive a practical privacy loss accounting method, and discuss its relation to the moments accountant. All the proofs are available in the supplementary material.

4.1. Definition

Let us define *strong* Bayesian differential privacy (Definition 1) and (*weak*) Bayesian differential privacy (Definition 2). The first provides a better intuition, connection to concentration inequalities, and is being used for privacy accounting. Unfortunately, it may not be closed under post-processing, and therefore, the actual guarantee provided by BDP is stated in Definition 2 and mimics the (ε, δ)-differential privacy (Dwork et al., 2014). The reason Definition 1 may pose a problem with post-processing is that it does not consider sets of outcomes, and a routine that integrates groups of values into one value could therefore invalidate the guarantee by increasing the probability ratio beyond epsilon. On the other hand, it can still be used for accounting with adaptive composition, because in this context, every next step is conditioned on a single outcome of the previous step. This separation mirrors the moments accountant approach of bounding tails of the privacy loss random variable and converting it to the (ε, δ)-DP guarantee (Abadi et al., 2016), but does so in a more explicit manner.

Definition 1 (Strong Bayesian Differential Privacy)**.** *A randomised function $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$, range $\mathcal{R}$, and outcome $w = \mathcal{A}(\cdot)$, satisfies $(\varepsilon_\mu, \delta_\mu)$-strong Bayesian differential privacy if for any two adjacent datasets $D, D' \in \mathcal{D}$,* differing in a single data point $x' \sim \mu(x)$*, the following holds:*

$$\Pr[L_{\mathcal{A}}(w, D, D') \geq \varepsilon_\mu] \leq \delta_\mu, \tag{2}$$

where probability is taken over the randomness of the outcome w and the additional example x'.

Here, $L_{\mathcal{A}}(w, D, D')$ is the privacy loss defined as

$$L_{\mathcal{A}}(w, D, D') = \log \frac{p(w|D)}{p(w|D')}, \tag{3}$$

where $p(w|D)$, $p(w|D')$ are private outcome distributions for corresponding datasets. For brevity, we often omit parameters and denote the privacy loss simply by L.

We use the subscript μ to underline the main difference between the classic DP and Bayesian DP: in the classic definition the probability is taken only over the randomness of the outcome (w), while the BDP definition contains two random variables (w and x'). Therefore, the privacy parameters ε and δ depend on the data distribution $\mu(x)$.

The addition of another random variable yields the change in the meaning of δ_μ compared to the δ of DP. In Bayesian differential privacy, it also accounts for the privacy mechanism failures in the tails of data distributions in addition to the tails of outcome distributions.

Definition 2 (Bayesian Differential Privacy)**.** *A randomised function $\mathcal{A} : \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies $(\varepsilon_\mu, \delta_\mu)$-Bayesian differential privacy if for any two adjacent datasets $D, D' \in \mathcal{D}$,* differing in a single data point $x' \sim \mu(x)$*, and for any set of outcomes $\mathcal{S}$ the following holds:*

$$\Pr\left[\mathcal{A}(D) \in \mathcal{S}\right] \leq e^{\varepsilon_\mu} \Pr\left[\mathcal{A}(D') \in \mathcal{S}\right] + \delta_\mu. \tag{4}$$

Proposition 1. *$(\varepsilon_\mu, \delta_\mu)$-strong Bayesian differential privacy implies $(\varepsilon_\mu, \delta_\mu)$-Bayesian differential privacy.*

[1] https://www.macobserver.com/analysis/google-apple-differential-privacy/

Bayesian DP repeats some basic properties of the classic DP, such as composition, post-processing resilience and group privacy. More details, proofs for these properties and the above proposition, can be found in supplementary material.

While Definitions 1 and 2 do not specify the distribution of any point in the dataset other than the additional example x', it is natural to assume that all examples in the dataset are drawn from the same distribution $\mu(x)$. This holds in many real-world applications, including applications evaluated in this paper, and it allows using dataset samples instead of requiring knowing the true distribution.

We also assume that data points are exchangeable (Aldous, 1985), i.e. any permutation of data points has the same joint probability. It enables tighter accounting for iterative applications of the privacy mechanism (see Section 4.2), is weaker than independence and is naturally satisfied in the considered scenarios.

4.2. Privacy Accounting

In the context of learning, it is important to be able to keep track of the privacy loss over iterative applications of the privacy mechanism. And since the bounds provided by the basic composition theorem are loose, we formulate the *advanced composition theorem* and develop a general accounting method for Bayesian differential privacy, the *Bayesian accountant*, that provides a tight bound on privacy loss and is straightforward to implement. We draw inspiration from the moments accountant (Abadi et al., 2016).

Observe that Eq. 4 is a typical concentration bound inequality, which are well studied in probability theory. One of the most common examples of such bounds is Markov's inequality. In its extended form, it states the following:

$$\Pr[|L| \geq \varepsilon_\mu] \leq \frac{\mathbb{E}[\varphi(|L|)]}{\varphi(\varepsilon_\mu)}, \tag{5}$$

where $\varphi(\cdot)$ is a monotonically increasing non-negative function. It is immediately evident that it provides a relation between ε_μ and δ_μ (i.e. $\delta_\mu = \mathbb{E}[\varphi(|L|)]/\varphi(\varepsilon_\mu)$), and in order to determine them we need to choose φ and compute the expectation $\mathbb{E}[\varphi(|L(w, D, D')|)]$. Note that $L(w, D, D') = -L(w, D', D)$, and since the inequality has to hold for any pair of D, D', we can use L instead of $|L|$.

We use the Chernoff bound that can be obtained by choosing $\varphi(L) = e^{\lambda L}$. It is widely known because of its tightness, and although not explicitly stated, it is also used by Abadi et al. (2016). The inequality in this case transforms to

$$\Pr[L \geq \varepsilon_\mu] \leq \frac{\mathbb{E}[e^{\lambda L}]}{e^{\lambda \varepsilon_\mu}}. \tag{6}$$

This inequality requires the knowledge of the moment generating function of L or some bound on it. The choice of the parameter λ can be arbitrary, because the bound holds for any value of it, but it determines how tight the bound is. By simple manipulations we obtain

$$\begin{aligned}\mathbb{E}[e^{\lambda L}] &= \mathbb{E}\left[e^{\lambda \log \frac{p(w|D)}{p(w|D')}}\right] \\ &= \mathbb{E}\left[\left(\frac{p(w|D)}{p(w|D')}\right)^{\lambda}\right].\end{aligned} \tag{7}$$

If the expectation is taken only over the outcome randomness, this expression is the function of Rényi divergence between $p(w|D)$ and $p(w|D')$, and following this path yields re-derivation of Rényi differential privacy (Mironov, 2017). However, by also taking the expectation over additional examples $x' \sim \mu(x)$, we can further tighten this bound.

By the law of total expectation,

$$\mathbb{E}\left[\left(\frac{p(w|D)}{p(w|D')}\right)^{\lambda}\right] = \mathbb{E}_x\left[\mathbb{E}_w\left[\left(\frac{p(w|D)}{p(w|D')}\right)^{\lambda}\middle| x'\right]\right], \tag{8}$$

where the inner expectation is again the function of Rényi divergence, and the outer expectation is over $\mu(x)$.

Combining Eq. 7 and 8 and plugging it in Eq. 6, we get

$$\Pr[L \geq \varepsilon_\mu] \leq \mathbb{E}_x\left[e^{\lambda \mathcal{D}_{\lambda+1}[p(w|D)\|p(w|D')] - \lambda \varepsilon_\mu}\right]. \tag{9}$$

This expression determines how to compute ε_μ for a fixed δ_μ (or vice versa) for one invocation of the privacy mechanism. However, to accommodate the iterative nature of learning, we need to deal with the composition of multiple applications of the mechanism. We already determined that our privacy notion is naively composable, but in order to achieve better bounds we need a tighter composition theorem.

Theorem 1 (Advanced Composition)**.** *Let a learning algorithm run for T iterations. Denote by $w^{(1)} \dots w^{(T)}$ a sequence of private learning outcomes at iterations $1, \dots, T$, and $L^{(1:T)}$ the corresponding total privacy loss. Then,*

$$\mathbb{E}\left[e^{\lambda L^{(1:T)}}\right] \leq \prod_{t=1}^{T} \mathbb{E}_x\left[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t\|q_t)}\right]^{\frac{1}{T}},$$

where $p_t = p(w^{(t)}|w^{(t-1)}, D)$, $q_t = p(w^{(t)}|w^{(t-1)}, D')$.

Proof. See supplementary material. □

Unlike the moments accountant, our composition theorem presents an upper bound on the total privacy loss due to computing expectation over the distribution of the same example over all iterations. However, we found that the inequality tends to be tight in practice, and there is little

overhead compared to naïvely swapping the product and the expectation.

We denote the logarithm of the quantity inside the product in Theorem 1 as $c_t(\lambda, T)$ and call it the *privacy cost* of the iteration t:

$$c_t(\lambda, T) = \log \mathbb{E}_x \left[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t \| q_t)} \right]^{\frac{1}{T}} \tag{10}$$

The privacy cost of the whole learning process is then a sum of the costs of each iteration. We can now relate ε and δ parameters of BDP through the privacy cost.

Theorem 2. *Let the algorithm produce a sequence of private learning outcomes $w^{(1)} \dots w^{(T)}$ using a known probability distribution $p(w^{(t)}|w^{(t-1)}, D)$. Then, for a fixed ε_μ:*

$$\log \delta_\mu \leq \sum_{t=1}^{T} c_t(\lambda, T) - \lambda \varepsilon_\mu.$$

Corollary 1. *Under the conditions above, for a fixed δ_μ:*

$$\varepsilon_\mu \leq \frac{1}{\lambda} \sum_{t=1}^{T} c_t(\lambda, T) - \frac{1}{\lambda} \log \delta_\mu.$$

Theorems 1, 2 and Corollary 1 immediately provide us with an efficient privacy accounting algorithm. During training, we compute the privacy cost $c_t(\lambda, T)$ for each iteration t, accumulate it, and then use to compute $\varepsilon_\mu, \delta_\mu$ pair. This process is ideologically close to that of the moment accountant but accumulates a different quantity (note the change from the privacy loss random variable to Rényi divergence). We further explore this connection in Section 4.4.2.

The link to Rényi divergence is an advantage for applicability of this framework: if the outcome distribution $p(w|D)$ has a known analytic expression for Rényi divergence (Gil et al., 2013; Van Erven & Harremos, 2014), it can be easily plugged into our accountant.

For the popular subsampled Gaussian mechanism (Abadi et al., 2016), we can demonstrate the following.

Theorem 3. *Given the Gaussian noise mechanism with the noise parameter σ and subsampling probability q, the privacy cost for $\lambda \in \mathbb{N}$ at iteration t can be expressed as*

$$c_t(\lambda, T) = \max\{c_t^L(\lambda, T), c_t^R(\lambda, T)\},$$

where

$$c_t^L(\lambda, T) = \frac{1}{T} \log \mathbb{E}_x \left[\mathbb{E}_{k \sim B(\lambda+1, q)} \left[e^{\frac{k^2 - k}{2\sigma^2} \|g_t - g_t'\|^2} \right]^T \right],$$

$$c_t^R(\lambda, T) = \frac{1}{T} \log \mathbb{E}_x \left[\mathbb{E}_{k \sim B(\lambda, q)} \left[e^{\frac{k^2 + k}{2\sigma^2} \|g_t - g_t'\|^2} \right]^T \right],$$

and $B(\lambda, q)$ is the binomial distribution with λ experiments and the probability of success q.

4.3. Privacy Cost Estimator

Computing $c_t(\lambda, T)$ precisely requires access to the data distribution $\mu(x)$, which is unrealistic. Therefore, we need an estimator for $\mathbb{E}[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t \| q_t)}]$.

Typically, having access to the distribution samples, one would use the law of large numbers and approximate the expectation with the sample mean. This estimator is unbiased and converges with the growing number of samples. However, these are not the properties we are looking for. The most important property of the estimator in our context is that it *does not underestimate* $\mathbb{E}[e^{T\lambda \mathcal{D}_{\lambda+1}(p_t \| q_t)}]$, because the bound (Eq. 6) would not hold for this estimate otherwise.

We employ the Bayesian view of the parameter estimation problem (Oliphant, 2006) and design an estimator with this single property: given a fixed γ, it returns the value that overestimates the true expectation with probability $1 - \gamma$. We then incorporate the estimator uncertainty γ in δ_μ.

4.3.1. Binary Case

Let us demonstrate the process of constructing the expectation estimator with the above property on a simple binary example. This technique is based on (Oliphant, 2006) and it translates directly to other classes of distributions with minor adjustments. Here, we also address the concern of not taking into account the data absent from the dataset.

Let the data $\{x_1, x_2, \dots, x_N\}$, such that $x_i \in \{0, 1\}$, have a common mean and a common variance. As this information is insufficient to solve our problem, let us also assume that the data comes from *the maximum entropy distribution*. This assumption adds the minimum amount of information to the problem and makes our estimate pessimistic.

For the binary data with the common mean ρ, the maximum entropy distribution is the Bernoulli distribution:

$$f(x_i|\rho) = \rho^{x_i}(1-\rho)^{1-x_i}, \tag{11}$$

where ρ is also the probability of success ($x_i = 1$). Then,

$$f(x_1, \dots, x_N|\rho) = \rho^{N_1}(1-\rho)^{N_0}, \tag{12}$$

where N_0 and N_1 is the number of 0's and 1's in the dataset.

We impose the flat prior on ρ, assuming all values in $[0, 1]$ are equally likely, and use Bayes' theorem to determine the distribution of ρ given the data:

$$f(\rho|x_1, \dots, x_N) = \frac{\Gamma(N_0 + N_1 + 2)}{\Gamma(N_0 + 1)\Gamma(N_1 + 1)} \rho^{N_1}(1-\rho)^{N_0}, \tag{13}$$

where the normalisation constant in front is obtained by setting the integral over ρ equal to 1.

Now, we can use the above distribution of ρ to design an estimator $\hat{\rho}$, such that it overestimates ρ with high probability,

i.e. $\Pr[\rho \leq \hat{\rho}] \geq 1-\gamma$. Namely, $\hat{\rho} = F^{-1}(1-\gamma)$, where F^{-1} is the inverse of the CDF:

$$F^{-1}(1-\gamma) = \inf\{z \in \mathbb{R} : \int_{-\infty}^{z} f(t|x_1,\ldots,x_N)dt \geq 1-\gamma\}.$$

We refer to γ as the *estimator failure probability*, and to $1-\gamma$ as the *estimator confidence*.

To demonstrate the resilience of this estimator to unseen data, consider the following example. Let the true expectation be 0.01, and let the data consist of 100 zeros, and no ones. A typical "frequentist" mean estimator would confidently output 0. However, our estimator would never output 0, unless the confidence is set to 0. When the confidence is set to 1 ($\gamma = 0$), the output is 1, which is the most pessimistic estimate. Finally, the output $\hat{\rho} = \rho = 0.01$ will be assigned the failure probability $\gamma = 0.99^{101} \approx 0.36$, which is the probability of not drawing a single 1 in 101 draws.

In a real-world system, the confidence would be set to a much higher level (in our experiments, we use $\gamma = 10^{-15}$), and the probability of 1 would be significantly overestimated. Thus, unseen data do not present a problem for this estimator, because it exaggerates the probability of data that increase the estimated expectation.

4.3.2. Continuous Case

For applications evaluated in this paper, we are primarily concerned with continuous case. Thus, let us define the following m-sample estimator of $c_t(\lambda, T)$ for continuous distributions with existing mean and variance:

$$\hat{c}_t(\lambda, T) = \log\left[M(t) + \frac{F^{-1}(1-\gamma, m-1)}{\sqrt{m-1}}S(t)\right], \tag{14}$$

where $M(t)$ and $S(t)$ are the sample mean and the sample standard deviation of $e^{\lambda \hat{D}^{(t)}_{\lambda+1}}$, $F^{-1}(1-\gamma, m-1)$ is the inverse of the Student's t-distribution CDF at $1-\gamma$ with $m-1$ degrees of freedom, and

$$\begin{aligned}\hat{D}^{(t)}_{\lambda+1} &= \max\left\{D_{\lambda+1}(\hat{p}_t \| \hat{q}_t),\ D_{\lambda+1}(\hat{q}_t \| \hat{p}_t)\right\},\\ \hat{p}_t &= p(w^{(t)} \mid w^{(t-1)}, B^{(t)}),\\ \hat{q}_t &= p(w^{(t)} \mid w^{(t-1)}, B^{(t)} \setminus \{x_i\}).\end{aligned} \tag{15}$$

Since in many cases learning is performed on mini-batches, we can similarly compute Rényi divergence on batches $B^{(t)}$.

Theorem 4. *Estimator $\hat{c}_t(\lambda, T)$ overestimates $c_t(\lambda, T)$ with probability $1-\gamma$. That is,*

$$\Pr\left[\hat{c}_t(\lambda, T) < c_t(\lambda, T)\right] \leq \gamma.$$

The proof follows the steps of the binary example above.

Remark. By adapting the maximum entropy probability distribution an equivalent estimator can be derived for other classes of distributions (e.g. discrete).

To avoid introducing new parameters in the privacy definition, we can incorporate the probability γ of underestimating the true expectation in δ_μ. We can re-write:

$$\begin{aligned}&\Pr[L_{\mathcal{A}}(w^{(t)}, D, D') \geq \varepsilon_\mu]\\ &\quad= \Pr\left[L_{\mathcal{A}}(w^{(t)}, D, D') \geq \varepsilon_\mu, \hat{c}_t(\lambda, T) \geq c_t(\lambda, T)\right]\\ &\qquad+ \Pr\left[L_{\mathcal{A}}(w^{(t)}, D, D') \geq \varepsilon_\mu, \hat{c}_t(\lambda, T) < c_t(\lambda, T)\right].\end{aligned} \tag{16}$$

When $\hat{c}_t(\lambda, T) \geq c_t(\lambda, T)$, using the Chernoff inequality, the first summand is bounded by $\beta = \exp(\sum_{t=1}^{T} \hat{c}_t(\lambda, T) - \lambda\varepsilon_\mu)$.

Whenever $\hat{c}_t(\lambda, T) < c_t(\lambda, T)$,

$$\begin{aligned}&\Pr[L_{\mathcal{A}}(w^{(t)}, D, D') \geq \varepsilon_\mu, \hat{c}_t(\lambda, T) < c_t(\lambda, T)]\\ &\quad\leq \Pr[\hat{c}_t(\lambda, T) < c_t(\lambda, T)]\\ &\quad\leq \gamma.\end{aligned} \tag{17}$$

Therefore, the true δ_μ is bounded by $\beta + \gamma$, and despite the incomplete data, we can claim that the mechanism is $(\varepsilon_\mu, \delta_\mu)$-Bayesian differentially private, where $\delta_\mu = \beta + \gamma$.

Remark. This step further changes the interpretation of δ_μ in Bayesian differential privacy compared to the classic δ of DP. Apart from the probability of the privacy loss exceeding ε_μ, e.g. in the tails of its distribution, it also incorporates our uncertainty about the true data distribution (in other words, the probability of underestimating the true expectation because of not observing enough data samples). It can be intuitively understood as accounting for unobserved (but feasible) data in δ_μ, rather than in ε_μ.

4.4. Discussion

4.4.1. Relation to DP

To better understand how the BDP bound relates to the traditional DP, consider the following conditional probability:

$$\Delta(\varepsilon, x') = \Pr\left[L(w, D, D') > \varepsilon \mid D, D' = D \cup \{x'\}\right]. \tag{18}$$

The moments accountant outputs δ that upper-bounds $\Delta(\varepsilon, x')$ for all x'. It is not true in general for other accounting methods, but let us focus on MA, as it is by far the most popular. Consequently, the MA bound is

$$\max_x \Delta(\varepsilon, x) \leq \delta, \tag{19}$$

where ε is a chosen constant. At the same time, BDP bounds the probability that is not conditioned on x', but we can transform one to another through marginalisation and get:

$$\mathbb{E}_x \left[\Delta(\varepsilon, x)\right] \leq \delta_\mu. \tag{20}$$

Since $\Delta(\cdot)$ is a non-negative random variable in x, we can apply Markov's inequality and obtain a tail bound on it using δ_μ. *We can therefore find a pair $(\varepsilon, \delta)_p$ that holds for any percentile p of the data distribution, not just in expectation.* In all our experiments, we consider bounds well above 99th percentile, so it is very unlikely to encounter data for which the equivalent DP guarantee doesn't hold. Moreover, it is possible to characterise privacy by building a curve for different percentiles, and hence, gain more insight into how well users and their data are protected.

4.4.2. Relation to Moments Accountant

As mentioned in Section 4.2, removing the distribution requirement on D, D' and further simplifying Eq. 9, we can recover the relation between Rényi DP and (ε, δ)-DP.

At the same time, our accounting technique closely resembles the moments accountant. In fact, we can show that the moments accountant is a special case of Theorem 3. Ignoring the data distribution information and substituting expectation by $\max_{D,D'}$ yields the substitution of $\|g_t - g'_t\|$ for C in Theorem 3, where C is the sensitivity (or clipping threshold), which turns out to be the exact moments accountant bound. In addition, there are some extra benefits, such as avoiding numerical integration when using $\lambda \in \mathbb{N}$ due to connection to Binomial distribution, which improves numerical stability and computational efficiency.

4.4.3. Privacy of $\hat{c}_t(\lambda, T)$

Due to calculating $\hat{c}_t(\lambda, T)$ from data, our privacy guarantee becomes data-dependent and may potentially leak information. To obtain a theoretical bound on this leakage, we need to get back to the maximum entropy assumption in Section 4.3, and assume that $M(t)$ and $S(t)$ are following some specific distributions, such as Gaussian and χ^2 distributions. Hence, in case of simple random sampling, these statistics for two neighbour datasets are differentially private and the privacy parameters can be computed using Rényi divergence. Furthermore, these guarantees are controlled by the number of samples used to compute the statistics: the more samples are used, the more accurate the statistics are, and the less privacy leakage occurs. This property can be used to control estimates privacy without sacrificing their tightness, only at the cost of extra computation time. Without distributional assumptions, the bound can be computed in the limit of the sample size used by the estimator, using the CLT.

On the other hand, consider the fact that the information from many high-dimensional vectors gets first compressed down to their pairwise distances, which are not as informative in high-dimensional spaces (i.e. the curse of dimensionality), and then down to one number. Intuitively, at this rate of compression very little knowledge can be gained by an attacker in practice.

The first approach would provide little information about real-world cases due to potentially unrealistic assumptions, and hence, we opt for the second approach. We examine pairwise gradient distances of the points within the training set and outside, and demonstrate that the privacy leakage is not statistically significant in practice (see Section 5.2).

5. Evaluation

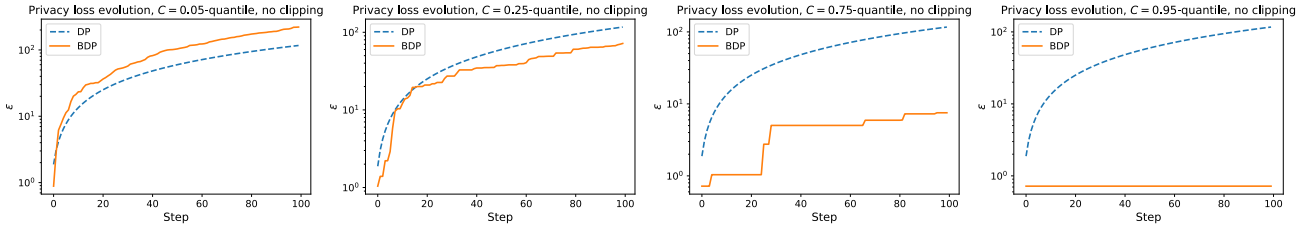
This experimental section comprises two parts. First, we examine how well Bayesian DP composes over multiple steps. We use the Bayesian accountant and compare to the state-of-the-art DP results obtained by the moments accountant. Second, we consider the context of machine learning. In particular, we use the differentially private stochastic gradient descent (DP-SGD), a well known privacy-preserving learning technique broadly used in combination with the moments accountant, to train neural networks on classic image classification tasks MNIST (LeCun et al., 1998) and CIFAR10 (Krizhevsky, 2009). We then compare the accuracy and privacy guarantees obtained under BDP and under DP. We also perform experiments with variational inference on Abalone (Waugh, 1995) and Adult (Kohavi, 1996) datasets.

Importantly, DP and BDP can use the same privacy mechanism and be accounted in parallel to ensure the DP guarantees hold if BDP assumptions fail. Thus, all comparisons in this section should be viewed in the following way: the reported BDP guarantee would apply to *typical* data (i.e. data drawn from the same distribution as the dataset); the reported DP guarantee would apply to all other data; their difference is the advantage for typical data we gain by using Bayesian DP. In some experiments we use smaller noise variance for BDP in order to speed up training, meaning that the reported BDP guarantees will further improve if noise variance is increased to DP levels. For more details and additional experiments, we refer the reader to the supplementary material, while the source code is available on GitHub[2].

5.1. Composition

First, we study the growth rate of the privacy loss over a number of mechanism invocations. This experiment is carried out using synthetic gradients drawn from the Weibull distribution with the shape parameter < 1 to imitate a more difficult case of heavy-tailed gradient distributions. We do

[2] https://github.com/AlekseiTriastcyn/bayesian-differential-privacy

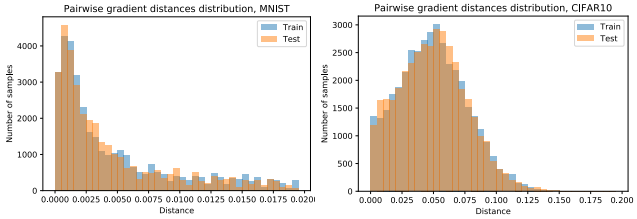


(a) 0.05-quantile of $\|\nabla f\|$. (b) 0.25-quantile of $\|\nabla f\|$. (c) 0.75-quantile of $\|\nabla f\|$. (d) 0.95-quantile of $\|\nabla f\|$.

Figure 1: Evolution of ε and ε_μ over multiple steps of the Gaussian noise mechanism with $\sigma = C$ for DP (with clipping) and BDP (without clipping). Sub-captions indicate the noise variance relative to the gradient norms distribution.

Table 1: Estimated privacy bounds ε for $\delta = 10^{-5}$ and $\delta_\mu = 10^{-10}$ for MNIST, CIFAR10, Abalone and Adult datasets. In parenthesis, a potential attack success probability $P(A)$.

	Accuracy		**Privacy**	
Dataset	Baseline	Private	DP	BDP
MNIST	99%	96%	2.2 (0.898)	**0.95 (0.721)**
CIFAR10	86%	73%	8.0 (0.999)	**0.76 (0.681)**
Abalone	77%	76%	7.6 (0.999)	**0.61 (0.649)**
Adult	81%	81%	0.5 (0.623)	**0.2 (0.55)**



(a) MNIST. (b) CIFAR10.

Figure 2: Pairwise gradient distances distribution.

not clip gradients for BDP in order to show the raw effect of the signal-to-noise ratio on the privacy loss behaviour. Technically, bounded sensitivity is not as essential for BDP, because extreme individual contributions are mitigated by their low probability. However, in practice it is still advantageous to have a better control over privacy loss spikes and ensure that the worst-case DP guarantee is preserved.

In Figure 1, we plot ε and ε_μ as a function of steps for different levels of noise. Naturally, as the noise standard deviation gets closer to the expected gradients norm, the growth rate of the privacy loss decreases dramatically. Even when the noise is at the 0.25-quantile, the Bayesian accountant matches the moments accountant. It is worth noting, that DP behaves the same in all these experiments because the gradients get clipped at the noise level C. Introducing clipping for BDP yields the behaviour of Figure 1d, as we demonstrate in the next section on real data.

5.2. Learning

We now consider the application to privacy-preserving deep learning. Our setting closely mimics that of (Abadi et al., 2016) to enable a direct comparison with the moments accountant and DP. We use a version of DP-SGD that has been extensively applied to build differentially private machine learning models. The idea of DP-SGD is to clip the gradient norm to some constant C (ensuring bounded sensitivity) and add Gaussian noise with variance $C^2\sigma^2$ at every iteration of SGD. For Abalone and Adult, we use variational inference in a setting similar to (Jälkö et al., 2016).

Using the gradient distribution information allows the BDP models to reach the same accuracy at a much lower ε (for 99.999% of data points from this distribution, see Section 4.4.1). On MNIST, we manage to reduce it from 2.2 to 0.95. For CIFAR10, from 8.0 to 0.76. See details in Table 1. Moreover, since less noise is required for Bayesian DP, the models reach the same test accuracy much faster. For example, our model reaches 96% accuracy within 50 epochs for MNIST, while DP model requires more noise and slower training over hundreds of epochs to avoid ε blowing up. These results confirm that discounting outliers in the privacy accounting process is highly beneficial for getting high accuracy and tighter guarantees for all the other points. To make our results more transparent, we include in Table 1 the potential attack success probability $P(A)$ computed using Eq. 1. In this interpretation, the benefits of using BDP become even more apparent.

An important aspect of BDP, discussed in Section 4.4.3, is the potential privacy leakage of the privacy cost estimator. To illustrate that this leakage is minimal, we conduct the following experiment. After training the model (to ensure it contains as much information about data as possible), we compute the gradient pairwise distances over train and test sets. We then plot the histograms of these distances to inspect any divergence that would distinguish the data that was used in training. Note that this is more information than what is available to an adversary, who only observes ε_μ, δ_μ.

As it turns out, these distributions are nearly identical (see Figures 2a and 2b), and we do not observe any correlation with the fact of the presence of data in the training set. For example, the sample mean of the test set can be both somewhat higher or lower than that of the train set. We also run the t-test for equality of means and Levene's test for equality of variances, obtaining p-values well over the 0.05 threshold, suggesting that the difference of the means and the variances of these distributions is not statistically significant and the equality hypothesis cannot be rejected.

6. Conclusion

We introduce the notion of $(\varepsilon_\mu, \delta_\mu)$-Bayesian differential privacy, a variation of (ε, δ)-differential privacy for sensitive data that are drawn from an arbitrary (and unknown) distribution $\mu(x)$. It is a reasonable assumption in many ML scenarios where models are designed for and trained on specific data distributions (e.g. emails, face images, ECGs, etc.). For example, trying to hide music records in a training set for ECG analysis may be unjustified, because the probability of it appearing is actually much smaller than δ.

We present the advanced composition theorem for Bayesian DP that allows for efficient and tight privacy accounting. Since the data distribution is unknown, we design an estimator that overestimates the privacy loss with high, controllable probability. Moreover, as the data sample is finite, we employ the Bayesian parameter estimation approach with the flat prior and the maximum entropy principle to avoid underestimating probabilities of unseen examples.

Our evaluation confirms that Bayesian DP is highly beneficial in ML context where its additional assumptions are naturally satisfied. First, it needs less noise for comparable privacy guarantees (with high probability, as per Section 4.4.1). Second, models train faster and can reach higher accuracy. Third, it may be used along with DP to ensure the worst-case guarantee for out-of-distribution samples and outliers, while providing tighter guarantees for most cases. In our supervised learning experiments, ε always remains below 1, translating to much more meaningful bounds on a potential attacker success probability.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pp. 308–318. ACM, 2016.

Aldous, D. J. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII1983*, pp. 1–198. Springer, 1985.

Bassily, R. and Freund, Y. Typical stability. *arXiv preprint arXiv:1604.03336*, 2016.

Bassily, R., Groce, A., Katz, J., and Smith, A. Coupled-worlds privacy: Exploiting adversarial uncertainty in statistical data privacy. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 439–448. IEEE, 2013.

Bhaskar, R., Bhowmick, A., Goyal, V., Laxman, S., and Thakurta, A. Noiseless database privacy. In *International Conference on the Theory and Application of Cryptology and Information Security*, pp. 215–232. Springer, 2011.

Blum, A., Ligett, K., and Roth, A. A learning theory approach to noninteractive database privacy. *Journal of the ACM (JACM)*, 60(2):12, 2013.

Bun, M. A teaser for differential privacy. 2017.

Bun, M. and Steinke, T. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer, 2016.

Duan, Y. Privacy without noise. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pp. 1517–1520. ACM, 2009.

Dwork, C. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052, pp. 1–12, Venice, Italy, July 2006. Springer Verlag. ISBN 3-540-35907-9. URL https://www.microsoft.com/en-us/research/publication/differential-privacy/.

Dwork, C. and Rothblum, G. N. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Dwork, C., McSherry, F., Nissim, K., and Smith, A. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006.

Dwork, C., Roth, A., et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Fredrikson, M., Jha, S., and Ristenpart, T. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333. ACM, 2015.

Gil, M., Alajaji, F., and Linder, T. Rényi divergence measures for commonly used univariate continuous distributions. *Information Sciences*, 249:124–131, 2013.

Hall, R., Rinaldo, A., and Wasserman, L. Random differential privacy. *arXiv preprint arXiv:1112.2680*, 2011.

He, X., Machanavajjhala, A., and Ding, B. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data*, pp. 1447–1458. ACM, 2014.

Hitaj, B., Ateniese, G., and Pérez-Cruz, F. Deep models under the gan: information leakage from collaborative deep learning. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 603–618. ACM, 2017.

Jälkö, J., Dikmen, O., and Honkela, A. Differentially private variational inference for non-conjugate models. *arXiv preprint arXiv:1610.08749*, 2016.

Kifer, D. and Machanavajjhala, A. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):3, 2014.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. Citeseer, 1996.

Krizhevsky, A. Learning multiple layers of features from tiny images. 2009.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Leung, S. and Lui, E. Bayesian mechanism design with efficiency, privacy, and approximate truthfulness. In *International Workshop on Internet and Network Economics*, pp. 58–71. Springer, 2012.

McMahan, H. B., Ramage, D., Talwar, K., and Zhang, L. Learning differentially private recurrent language models. 2018.

Mironov, I. Renyi differential privacy. In *Computer Security Foundations Symposium (CSF), 2017 IEEE 30th*, pp. 263–275. IEEE, 2017.

Oliphant, T. E. A bayesian perspective on estimating mean, variance, and standard-deviation from data. 2006.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., and Talwar, K. Semi-supervised knowledge transfer for deep learning from private training data. *arXiv preprint arXiv:1610.05755*, 2016.

Papernot, N., Song, S., Mironov, I., Raghunathan, A., Talwar, K., and Erlingsson, Ú. Scalable private learning with pate. *arXiv preprint arXiv:1802.08908*, 2018.

Shokri, R. and Shmatikov, V. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pp. 1310–1321. ACM, 2015.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pp. 3–18. IEEE, 2017.

Van Erven, T. and Harremos, P. Rényi divergence and kullback-leibler divergence. *IEEE Transactions on Information Theory*, 60(7):3797–3820, 2014.

Waugh, S. G. *Extending and benchmarking Cascade-Correlation: extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks*. PhD thesis, University of Tasmania, 1995.

Yang, B., Sato, I., and Nakagawa, H. Bayesian differential privacy on correlated data. In *Proceedings of the 2015 ACM SIGMOD international conference on Management of Data*, pp. 747–762. ACM, 2015.