Abstract

We consider the problem of differential privacy accounting, i.e., estimation of privacy loss bounds, in machine learning in a broad sense. We propose two versions of a generic privacy accountant suitable for a wide range of learning algorithms. Both versions are derived in a simple and principled way using well-known tools from probability theory, such as concentration inequalities. We demonstrate that our privacy accountant is able to achieve state-of-the-art estimates of DP guarantees and can be applied to new areas like variational inference. Moreover, we show that the latter enjoys differential privacy at minor cost.

1. Introduction

The rise of data analytics and machine learning (ML) presents countless opportunities for companies, governments and individuals to benefit from the accumulated data. At the same time, their ability to capture fine levels of detail potentially compromises 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 argue about the presence of individual records in the training set or recover certain features of these records. To tackle this problem a number of solutions has been proposed. They vary in how privacy is achieved and to what extent data is protected. In this paper, we consider models that achieve differential privacy (DP) guarantees (Dwork, 2006), as it is widely accepted by researchers and practitioners as a rigorous standard. Initially, DP algorithms focused on sanitising simple statistics, such as mean, median, etc., using a technique known as output perturbation. In recent years, the field made a lot of progress towards the goal of privacy-preserving machine learning, through works on objective perturbation (Chaudhuri et al., 2011), SGD with DP updates (Song et al., 2013), to more complex and practical techniques (Abadi et al., 2016; Papernot et al., 2016; 2018; McMahan et al., 2018).

We investigate a specific aspect of the problem: privacy accounting. In general, it means keeping track of privacy loss accumulating during repetitive application of privacy mechanisms, and transforming it into DP parameters $\varepsilon$ and $\delta$. The majority of earlier work uses already existing accounting methods, focusing on developing and improving privacy-preserving ML algorithms, frameworks, or applying existing techniques to new areas and models. The simplest examples of accounting are basic and advanced composition theorems (Dwork et al., 2014). Most notably, the issue of better privacy accounting has been raised by Abadi et al. (2016) who presented the moments accountant (MA), a state-of-the-art method currently used by most of the latest work in the area (Park et al., 2016a;b; Papernot et al., 2016; 2018; Jälfö et al., 2016; Geyer et al., 2017). While significantly improving tightness of the privacy bounds (bounds on $\varepsilon$ and $\delta$), this technique has certain disadvantages. For instance, it is developed specifically for the Gaussian noise mechanism and requires adaptation to be used in other contexts. Moreover, as we demonstrate below, the bounds can be significantly improved.

Our goal is to develop a more generic framework for privacy accounting. Particularly, we are interested in pushing the current state-of-the-art in two directions: computing tighter bounds on $(\varepsilon, \delta)$-DP parameters and ensuring broad out-of-the-box applicability to common machine learning methods.

The main idea of our accounting technique is to consider privacy loss as a random variable and use principled tools from probability theory to find concentration bounds on it. While it is similar to the moments accountant, we take this approach further and demonstrate that it considerably improves upon MA bounds. The final result is a simple to implement, effective, and easily understandable framework for privacy accounting without deviating to alternative privacy notions.

Our contributions in this paper are the following:

- we propose a novel, principled approach for privacy accounting in learning with tighter privacy bounds;
- we further improve bound estimates by using a
Bayesian approach and forming a data-dependent belief about it;

- we experimentally demonstrate advantages of our method, including the state-of-the-art bound tightness and applicability to variational inference, a popular class of algorithms rarely considered in privacy research.

The remainder of the paper is structured as follows. Section 2 provides an overview of related work. Preliminaries and setting description is given in Section 3. In Sections 4 and 5, we derive and describe two versions of our privacy accountant and discuss related questions. Section 6 contains experimental results, and Section 7 concludes the paper.

2. Related Work

As machine learning applications become more and more common, various vulnerabilities and attacks on ML models get discovered (for example, model inversion (Fredrikson et al., 2015) and membership inference (Shokri et al., 2017)), raising the need of developing matching defences.

Differential privacy (Dwork, 2006; Dwork et al., 2006) is one of the strongest privacy standards that can be employed to protect ML models from these and other attacks. Since pure DP is hard to achieve in many realistic complex learning tasks, a notion of approximate $(\epsilon, \delta)$-DP is used across-the-board in machine learning. It is often achieved as a result of applying the Gaussian noise mechanism (Dwork et al., 2014). Lately, several alternative notions and relaxations of DP have been proposed, such as concentrated DP (CDP) (Dwork & Rothblum, 2016; Bun & Steinke, 2016; Bun et al., 2018) and Rényi DP (RDP) (Mironov, 2017), allowing for easier privacy analysis.

Privacy analysis in the context of differentially private ML is often done post hoc and lies in finding parameters $\epsilon, \delta$ (or bounds on it) that apply to the entire learning process, as opposed to fixing $\epsilon, \delta$ beforehand and calibrating noise to satisfy it. Due to the nature of such analysis (keeping track and accumulating some quantity representing privacy loss during training) it is referred to as accounting. The simplest accounting can be done by using basic and advanced composition theorems (Dwork et al., 2014). However, bounds on $\epsilon, \delta$ obtained this way are prohibitively loose: using basic composition for big neural networks, $\epsilon$ can be in order of millions, so the DP guarantee loses any meaning.

Despite significantly improving bounds tightness, advanced composition is still not sufficient to get meaningful bounds for deep learning. This question has been considered by Abadi et al. (2016), and resulted in a privacy accounting method, called moments accountant, that is currently used by majority of the research community in this area. Just a small fraction of papers using moments accountant includes (Park et al., 2016a;b; Papernot et al., 2016; 2018; Geyer et al., 2017).

Apart from sharp bounds, moments accountant is attractive because it operates within the classical notion of $(\epsilon, \delta)$-DP. Alternative notions of DP also provide tight composition theorems, along with some other advantages, but to the best of our knowledge, are not broadly used in practice compared to traditional DP (although there are some examples (Geumlek et al., 2017)). One of the possible reasons for that is interpretability: parameters of $(\alpha, \epsilon)$-RDP or $(\mu, \tau)$-CDP are hard to interpret and hard to explain to a person without significant background in the area. At the same time, $\epsilon$ and $\delta$ can be easily understood intuitively, even though with some simplifications. Our goal in this work is to remain within a well-understood concept of $(\epsilon, \delta)$-DP, operate in a simple way similar to moments accountant, but improve the sharpness of its bounds and extend it to a broader range of privacy mechanisms within machine learning context.

We evaluate our method on two popular classes of learning algorithms: deep neural networks and variational inference (VI). Privacy-preserving deep learning is now extensively studied, and is frequently used in combination with moments accountant (Abadi et al., 2016; Papernot et al., 2016; 2018), which makes it a perfect setting for comparison. Bayesian inference methods, on the other hand, receive less attention from the private learning community. There are, nonetheless, very interesting results suggesting one could obtain DP guarantees “for free” (without adding noise) in some methods like posterior sampling (Dimitrakakis et al., 2014; 2017) and stochastic gradient Monte Carlo (Wang et al., 2015). A differentially private version of variational inference, obtained by applying noise to the gradients and using moments accountant, has also been proposed (Jälkä et al., 2016). We show that with our accountant it is possible to build VI that is both highly accurate and differentially private by sampling from variational distribution.

3. Problem Statement

In this section, we state our notation and provide some definitions used in the paper. We also describe a general setting of the problem.

3.1. Definitions and notation

We use $d, d'$ to represent neighbouring (adjacent) datasets. If not specified, it is assumed that these datasets differ in a single example. Individual examples in a dataset are denoted by $x, x'$, and learning outcomes (model parameters, neural network weights, etc.) by $w$.

Definition 1. A randomised function (algorithm) $A : \mathcal{D} \rightarrow \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies $(\epsilon, \delta)$-differential...
We assume a general iterative learning process, such that each iteration $t$ produces a (private) learning outcome $w^{(t)}$ that is used as a starting point for the next iteration. This learning outcome can be made private by applying some privacy mechanism (e.g., a Gaussian noise mechanism) or by drawing it from a distribution. In both cases, we say it comes from $p(w^{(t)}|w^{(t-1)}, d)$ (here we assumed a Markov property of the learning process, but it is not necessary in general). The process can run on subsamples of data, in which case $w^{(t)}$ comes from the distribution $p(w^{(t)}|w^{(t-1)}, b^{(t)})$, where $b^{(t)}$ is a batch of data used for parameters update in iteration $t$, and privacy is amplified through sampling (Balle et al., 2018). For each iteration we compute a quantity $c_t$ (we call it a privacy cost) that accumulates over the learning process and allows to compute $\varepsilon, \delta$ parameters of DP using concentration inequalities.

The overall privacy accounting workflow we suggest does not significantly differ from prior work. The main distinction is that it is not tied to a specific algorithm or a class of algorithms, as long as one can map it to the above setting.

In the two following sections, we define two ways of calculating the privacy cost $c_t$ and corresponding expressions that link it to $\varepsilon$ and $\delta$.

### 4. Generic Privacy Accountant

In this section, we introduce the generic accountant (GA) that uses the knowledge of the outcome distribution. There are no additional assumptions, and thus, it can be applied in all situations where the moments accountant is used, at the same time providing provably better bounds.

#### 4.1. Deriving privacy accountant

We approach the problem of estimating differential privacy parameters $\varepsilon, \delta$ in the following way. First, Definition 1 can be reformulated (see for example Appendix A of (Dwork et al., 2014)) as:

$$\Pr[L(w, d, d') \geq \varepsilon] \leq \delta. \quad (2)$$

At the same time, this expression clearly resembles typical concentration bounds well-known in probability theory, such as the following extension of Markov’s inequality:

$$\Pr[|L| \geq \varepsilon] \leq \mathbb{E}[\varphi(|L|)] / \varphi(\varepsilon), \quad (3)$$

where $\varphi(\cdot)$ is a monotonically increasing non-negative function. Thus, in order to determine $\varepsilon$ and $\delta$ (the r.h.s. of the inequality), we need to choose $\varphi$ and compute the expectation $\mathbb{E}[\varphi(|L(w, d, d')|)]$. First, we are going to do it only over $w$ while taking a maximum over $d, d'$ (similarly to MA). Then, in Section 5, we are going to naturally extend the method by computing an expectation over $w, x, x'$ and obtain tighter data-dependent bounds under mild additional assumptions.
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

\[
\Pr[L \geq \varepsilon] \leq \frac{E[e^{\lambda L}]}{e^{\lambda \varepsilon}}. \tag{4}
\]

Evidently, this inequality requires the knowledge of the moment generating function of \( L \) (or some bound on it), like the one used by moments accountant) and the choice of parameter \( \lambda \). By simple manipulations and observing similarities with R\'enyi divergence, we obtain

\[
E_{w,d,d'}[e^{\lambda L}] \leq \max_{d,d'} E_w \left[ e^{\lambda \log \frac{p(w|d)}{p(w|d')}} \right]
= \max_{d,d'} E_w \left[ \left( \frac{p(w|d)}{p(w|d')} \right)^{\lambda} \right]
= \max_{d,d'} e^{\lambda D_{\lambda+1}[p(w|d)||p(w|d')]}.
\tag{5}
\]

And plugging the above in Eq. 4,

\[
\Pr[L \geq \varepsilon] \leq \max_{d,d'} e^{\lambda D_{\lambda+1}[p(w|d)||p(w|d')]} - \lambda \varepsilon. \tag{6}
\]

This expression determines how to compute \( \varepsilon \) for a fixed \( \delta \) (or vice versa) for one step 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.

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

\[
E \left[ e^{\lambda L^{(1:T)}} \right] = e^{\lambda \sum_{t=1}^{T} D_{\lambda+1}(p_t||q_t)},
\]

where \( p_t = p(w^{(t)}|w^{(t-1)}, \ldots, w^{(1)}, d) \) and \( q_t = p(w^{(t)}|w^{(t-1)}, \ldots, w^{(1)}, d') \).

The proof of this theorem can be obtained by calculations analogous to Proof of Theorem 2.1 in (Abadi et al., 2016).

We can now combine these results and state the theorem used as a foundation for our accountant.

**Theorem 2.** Let the algorithm produce a sequence of private learning outcomes \( w^{(1)} \ldots w^{(T)} \) using a known distribution \( p(w^{(t)}|w^{(t-1)}, \ldots, w^{(1)}, d) \). Then, \( \varepsilon \) and \( \delta \) are related as

\[
\delta = \min_{\lambda} \max_{d,d'} e^{\sum_{t=1}^{T} c_t(\lambda) - \lambda \varepsilon},
\]

where \( c_t(\lambda) = \lambda D_{\lambda+1}(p_t||q_t). \)

As stated above, to account privacy one needs to keep track of the quantity \( c_t(\lambda) \) at each iteration and then translate it to \( \varepsilon, \delta \) pair by fixing one of them. The link to R\'enyi divergence allows computing \( c_t(\lambda) \) analytically for many common outcome distributions \( p(w|d) \) (Gil et al., 2013; Van Erven & Harremos, 2014).

**Optimal choice of \( \lambda \).** Chernoff inequality holds for any parameter \( \lambda > 0 \), and thus, the optimal estimates for \( \varepsilon, \delta \) one should minimise the r.h.s. over \( \lambda \), as stated in Theorem 2. While Abadi et al. (2016) suggest computing moments for \( \lambda \leq 32 \), we observe that since the moment generating function is log-convex it is possible to find an optimal value of \( \lambda \) that minimises the total bound. For some distributions, as shown below for Gaussian, it can be found analytically by computing the derivative and setting it to 0. In Section 6, we show how \( \varepsilon \) depends on the choice of \( \lambda \).

### 4.2. Gaussian mechanism

Consider the Gaussian noise mechanism (Dwork et al., 2014). The outcome distribution \( p(w|d) \) in this case is a Gaussian with known parameters. It is easy to see that the bound obtained by the generic accountant is tighter than that of the moments accountant, because it uses the exact expression for the r.h.s. of Chernoff inequality while the moments accountant uses an upper bound or numerical integration. More specifically, using expressions for R\'enyi divergence between two Gaussian distributions and clipping gradient norms to 1, we get

\[
\Pr[L \geq \varepsilon] \leq \min_{\lambda} e^{\frac{\lambda^2}{\varepsilon q^2} - \lambda \varepsilon}, \tag{7}
\]

where \( q \) is the batch sampling probability. This bound is tighter that the corresponding result in Appendix B of (Abadi et al., 2016). Note also that our bound does not require any assumptions on \( \lambda, q, \) or \( \sigma \).

We can further improve the bound by choosing \( \lambda \) optimally. Setting derivative of the above to 0, we get

\[
\lambda^*(\varepsilon) = \frac{\varepsilon \sigma^2}{q^2} - \frac{1}{2}. \tag{8}
\]

### 4.3. Discussion

As stated above, the generic accountant can be used in all contexts where the moment accountant is applicable. Moreover, it is easier to adapt to common continuous outcome distributions, including Gaussians with different variances, because of existing analytic expressions for R\'enyi divergence. In particular, it enables performing privacy accounting in VI (see Section 6.2) without adding noise.

**Privacy amplification.** Privacy amplification through sampling, as seen in Eq. 7 and in moments accountant, works
well in the case of Gaussian distributions. Plugging the sampling probability \( q \) in the formula for Rényi divergence reduces the bound by a factor of \( q^2 \).

It may not be as strong in general, for other probability distributions. However, there are general results for privacy amplification for Rényi DP that can be applied to our accountant, e.g. (Balle et al., 2018). We leave this investigation for future work.

**Relation to RDP and CDP.** One may notice connections to Rényi differential privacy (Mironov, 2017), which also uses Rényi divergence. Indeed, from Eq. 6 we can recover a relation between \( (\varepsilon, \delta) \)-DP and RDP. Moreover, considering subgaussian random variables we can also arrive to concentrated differential privacy (Dwork & Rothblum, 2016) bounds. We regard these connections as an additional confirmation of our findings.

### 5. Bayesian Privacy Accountant

In the previous section, we bounded the total expectation \( \mathbb{E}_{w,d,d'} \) by the maximum over datasets \( \max_{d,d'} \mathbb{E}_w \) (see Eq. 5). However, since machine learning normally operates on specific data, many pairs \( d \) and \( d' \) are highly unlikely, and these extreme values can make \( \max_{d,d'} \) much greater than \( \mathbb{E}_{d,d'} \).

Assuming access to the joint distribution of all datasets \( p(d, d') \) is unrealistic. However, if we change to the local privacy model, we can work with the expectation \( \mathbb{E}_{w,x,x'} \), which is more convenient because \( d \) is supposed to comprise samples \( x \sim p(x) \). If we additionally require \( x' \sim p(x) \), we can obtain a better (data-dependent) bound on privacy loss.

We call this version of the method the Bayesian accountant (BA) because of the different perspective on the problem. Rather than trying to find or bound some true \( \varepsilon \) and \( \delta \), it forms a belief about these values based on available data. In this case, naturally, \( \varepsilon \) and \( \delta \) become data-dependent and therefore should not be revealed directly. These values can be sanitised prior to publishing or just used for internal privacy audit, mechanism comparisons and research as a better alternative to absolute worst-case bounds.

The case of data-dependent \( \varepsilon, \delta \) is not novel. For example, choosing clipping thresholds in (Abadi et al., 2016) based on true gradients (e.g. using median) makes estimates data-dependent. Similar dependencies are present in (Geyer et al., 2017). Papernot et al. (2016) discuss this issue and propose using smooth sensitivity (Nissim et al., 2007) to sanitise their estimates. We leave this matter for future work.

#### 5.1. Estimating privacy loss from data

Let us start by deriving an estimator for \( \mathbb{E}_{w,x,x'}[e^{\lambda L(t)}] \). Then, since the bound (Eq. 4) only holds for the true expectation value, we will take into account the estimation error by incorporating it in \( \delta \) (in Section 5.2).

The following composition theorem is nearly identical to Theorem 1. However, in order to avoid computing expectation over \( x, x' \) for all \( T \) iterations at once, which would require re-running learning process many times, we have to extend our assumption: we suppose datasets \( d, d' \) consist of i.i.d. samples \( x \sim p(x) \). This assumption is already very common for the machine learning context we target. Moreover, it is also not novel in local privacy research (see, for example, (Kairouz et al., 2014)).

**Theorem 3.** Let the learning algorithm run for \( T \) iterations. Denote by \( w^{(1)} \ldots w^{(T)} \) a sequence of private learning outcomes obtained at iterations 1, \ldots, \( T \), and \( L^{(1:T)} \) the corresponding total privacy loss. Then,

\[
\mathbb{E}_{w,x,x'}[e^{\lambda L}] = \prod_{t=1}^{T} \mathbb{E}_{x,x'}[e^{\lambda L_{x,x'}^{t+1}(p_t||q_t)}],
\]

where \( q_t = p(w^{(t)}|w^{(t-1)}), \ldots, w^{(1)},x \) and \( p_t = p(w^{(t)}|w^{(t-1)}), \ldots, w^{(1)},x' \).

Proof of this theorem follows the same calculations as Theorem 1. As mentioned above, we move the product outside the exponent to avoid re-running the entire learning when computing the expectation over \( x, x' \). This is possible due to assumed independence of examples in \( d \).

The overall accounting workflow remains the same, with a differently defined privacy cost.

**Theorem 4.** Let the algorithm produce a sequence of private learning outcomes \( w^{(1)} \ldots w^{(T)} \) using a known distribution \( p(w^{(t)}|w^{(t-1)}, \ldots, w^{(1)}, d) \). Then, \( \varepsilon \) and \( \delta \) are related as

\[
\delta = \min_{\lambda} \sum_{t=1}^{T} c_t(\lambda) - \lambda \varepsilon,
\]

where \( c_t(\lambda) = \log \frac{1}{|B|} \sum_{x,x' \in B} e^{\lambda D_{x,x'}(p_t||q_t)} \) and \( B \) is a random subsample of the dataset picked for estimation.

Here we use the law of large numbers and estimate \( \mathbb{E}_{x,x'}[\cdot] \) by sampling pairs \( x, x' \) from the dataset and computing the corresponding average.

Given that \( \max_{d,d'} \) in Theorem 2 would normally be incorporated in \( c_t(\lambda) \), there is no difference in the final bound computation code between these two accountants. The only difference is how \( c_t(\lambda) \) is calculated in each learning iteration.

#### 5.2. Estimator uncertainty

Finally, Theorem 4 with the current definition of \( c_t(\lambda) \) only holds asymptotically (as \( |B| \to \infty \)). To take into account finite sample sizes we will add another uncertainty term to
δ. It is not going to change anything conceptually, except that δ will additionally incorporate the probability of the estimate being below the actual expectation.

Following the proof of Theorem 2.2 by Abadi et al. (2016), we can re-write the definition of DP:

\[
\Pr[A(d) \in S] = \Pr[A(d) \in S \cap B] + \Pr[A(d) \in S \cap B^c] 
\leq e^\epsilon \Pr[A(d') \in S] + \Pr[A(d) \in B],
\]

where \( B = \{ w : L(w, d, d') \geq \epsilon \} \) and \( S \) is any set of outcomes.

Now let \( L = e^{\sum_{t=0}^{T} c_t(\lambda)} \) and \( F = \{ w : \hat{L} + \kappa < \mathbb{E}[e^{\lambda L}] \} \), for some \( \kappa > 0 \), and denote \( \delta' = \Pr[A(d) \in B \cap F] \). Then,

\[
\Pr[A(d) \in B] \leq \delta + \delta' = \delta'',
\]

where \( \delta = \Pr[A(d) \in B \cap F^c] \) and is calculated by Theorem 4 using \( \hat{L} + \kappa \) as an estimate of \( \mathbb{E}[e^{\lambda L}] \). Therefore, we arrive at the definition of \( (\epsilon, \delta'') \)-DP:

\[
\Pr[A(d) \in S] \leq e^\epsilon \Pr[A(d') \in S] + \delta''.
\]

Depending on whether we fix \( \epsilon \) or \( \delta \), we have to find either \( \delta' \) or \( \kappa \). It can be done by employing a Bayesian view of the problem (Oliphant, 2006). More specifically, without assuming anything about \( L \), other than that it has a mean and a variance, we can apply a maximum entropy principle using an uninformative (flat) prior and show that the quantity \( \frac{L - \mathbb{E}[e^{\lambda L}]}{\sqrt{n - 1}} \) is distributed as Student’s t-distribution with \( n - 1 \) degrees of freedom, where \( S^2 \) is the sample variance and \( n \) is the number of samples (Oliphant, 2006). Then, \( \kappa \) or \( \delta' \) can be computed from the Student’s t CDF.

6. Evaluation

This experimental section is based on application of privacy accounting in two learning methods. The first one, differentially private stochastic gradient descent, is a well known privacy-preserving learning technique and is broadly used in combination with moments accountant (Abadi et al., 2016). The second one, variational inference, is largely overlooked in private learning research despite being a popular probabilistic method.

To evaluate our methods we train differentially private models using corresponding learning algorithms up to a certain level of accuracy while estimating privacy bounds by three accountants: moments accountant (MA), generic accountant (GA) and Bayesian accountant (BA). We then compare the tightness of the bounds, their dependence on hyper-parameters, and the privacy-accuracy trade-off.

Note that we have already shown theoretically that our bounds on \( \epsilon \) and \( \delta \) are sharper than that of the moments accountant. The goal of these experiments is to see what are the benefits in realistic learning situations and how hyper-parameters influence accountants’ estimates.

6.1. Differentially Private SGD

We first consider differentially private SGD (Abadi et al., 2016), which enables direct comparison with moments accountant (Abadi et al., 2016) in an environment for which it was developed. This method has been extensively applied to build differentially private machine learning models, from deep neural networks to Bayesian learning. The idea of DP-SGD is simple: at every iteration of SGD, clip the gradient norm to some constant \( C \) (ensuring bounded function sensitivity), and then add Gaussian noise with variance \( C^2\sigma^2 \). To account the total privacy loss, Abadi et al. (2016) accumulate bounds on moments and then transform it to \( \epsilon, \delta \).

We seek comparison in a setting very similar to the one described in (Abadi et al., 2016). We train a classifier represented by a neural network (without PCA) on MNIST dataset (LeCun et al., 1998) using DP-SGD. The dataset contains 60,000 training examples and 10,000 testing images. We use small batch sizes of 64, and clip gradient norms to \( C = 1 \). Unless otherwise specified, we fix \( \delta = 10^{-5} \).

Applying our accountants in this setting is simple. The learning outcome at each iteration is the gradient of the loss function w.r.t. model parameters, the outcome distribution is the noise distribution (originally a Gaussian, but in our case any distribution with a known expression for Rényi divergence). For generic accounting, the privacy cost \( c_t(\lambda) \) is calculated at each iteration using the clipping bound \( C \) and the noise parameter \( \sigma \). The cost is accumulated and transformed to \( \epsilon \) at the end of a fixed number of iterations for a pre-defined \( \delta \). For Bayesian accounting, pairs of examples
An additional advantage of the generic accountant is computational efficiency: on our machine, MA took around 30s to complete 1000 iterations of accounting, while GA completed it in $\sim 0.02s$. This notable difference is due to the fact that our method uses simple analytical expressions as opposed to numerical integration.

### 6.2. Variational Inference

While DP-SGD is widely applicable, some machine learning and statistical inference techniques do not require additional noise at all. For example, it has been shown that differential privacy guarantees arise naturally and "for free" in methods like sampling from the true posterior (Dimitrakakis et al., 2014) and Stochastic Gradient MCMC (Wang et al., 2015). Using our privacy accounting we can show that another popular Bayesian approach—variational inference—also enjoys almost "free" privacy guarantees.

The goal of variational inference is to approximate a posterior distribution $p(w|d)$ by a member of a known family of "simple" distributions $q(w;\theta)$ parametrised by $\theta$. Most commonly, it is done via minimising the reverse KL-divergence $D_{KL}(Q||P)$, but there are a lot of modern variations, for example using $\chi$-divergence (Dieng et al., 2017), Rényi divergence (Li & Turner, 2016), or other variational bounds (Chen et al., 2018).

As baselines, we use DPVI-MA (Jälkö et al., 2016) and DP-SGLD (Wang et al., 2015). The first one employs DP-SGD combined with moments accountant to train private VI model, while the second is a stochastic gradient MCMC method achieving DP due to the noisy nature of SGLD algorithm. Following (Jälkö et al., 2016), we run evaluation on two classification tasks taken from UCI database: Abalone (Waugh, 1995) and Adult (Kohavi, 1996). Both

| Dataset  | Accuracy | MA   | GA   | BA   |
|----------|----------|------|------|------|
| MNIST    | 96%      | 2.18 | 1.62 | 0.95 |
| Abalone  | 76%      | 7.6  | 4.54 | 0.11 |
| Adult    | 81%      | 2.90 | 2.70 | 0.08 |

$x_i, x_j$ are sampled from the dataset at each iteration and are used, along with $\sigma$, to compute $c_i(\lambda)$. Note that with the Bayesian accountant clipping gradients is no longer necessary, but it can still be done to ensure DP bounds for cases when assumptions of Section 5 are not satisfied (e.g. data samples are not i.i.d.).

Depicted in Figure 1 is the evolution of $\varepsilon$ bound over training epochs computed using the three methods. We observe that at all times both the generic accountant and the Bayesian accountant remain below the moments accountant curve, as predicted by theory. Similarly, the Bayesian accountant using more available information is always tighter than the generic accountant. By the end of 100 epochs, $\varepsilon$ estimated by Bayesian accountant does not exceed 1. It means that in the worst-case, probability of some learning outcome can change by a factor of at most 2.5, while according to moments accountant this factor is $\sim 9$. See details in Table 1. Another perspective is presented in Figure 2: for the same privacy loss bound $\varepsilon$, the failure probability obtained by our GA is $10^5$ times smaller than the one obtained by MA.

Figure 3 shows dependencies between $\varepsilon$ bounds, the choice of $\lambda$ and duration of learning. The optimal value of $\lambda$ can vary considerably depending on learning hyper-parameters, and in some cases, the quality of bounds can strongly depend on it (e.g. for many iterations, large batch sizes, etc.).
are binary classification tasks: predicting the age of abalone from physical measurements, and predicting income based on person’s attributes. They have 4,177 and 48,842 examples with 8 and 14 attributes accordingly. We use the same pre-processing and models as (Jalkó et al., 2016).

To translate VI to the language of our privacy accountant, $q(w; \theta)$ is the outcome distribution, and we are interested in

$$\mathbb{E} \left[ e^{\lambda \log \frac{q(w; \theta')}{q(w; \theta)}} \right],$$

where $\theta, \theta'$ were learnt from $d$ and $d'$. At each learning iteration, $w^{(t)}$ are sampled from $q(w; \theta^{(t-1)})$, updates are computed using the variational bound and data $d$ (or its subsamples), and parameters are updated to $\theta^{(t)}$. Therefore, for the generic accountant we need to bound the change of $\theta$ (to compute $\max_{d,d'}$ in Theorem 2), while for the Bayesian accountant we sample $x, x'$ and $w$ from $d$ and $q(w; \theta^{(t-1)})$ and estimate $c_{\epsilon}(\lambda)$ as it’s defined in Theorem 4.

Analogously to DP-SGD, privacy cost is accumulated during training and is transformed to $\varepsilon, \delta$ at the end of training.

To enable differential privacy for variational inference methods, we have to deal with two important restrictions. First, parameters $\theta$ of variational distribution $q(w; \theta)$ are not differentially private and need to be concealed or made private by other means. Second, as a result of the previous point, MAP or MLE estimates based on $\theta$ would also reveal private information. However, samples $w \sim q(w; \theta)$ are differentially private and can be used to perform the same tasks. We haven’t observed significant loss of accuracy when using a batch of samples $w$ instead of true parameters $\theta$, and thus, we consider it a minor cost. Note also that each sample needs to be accounted for, both in training and after training. In our tests, we run logistic regression using an average of 10 samples from variational distribution.

We observe in Figures 4 and 5 that our modified variational inference with Bayesian accountant achieves a strong advantage over DPVI-MA and DP-SGLD both in terms of accuracy and privacy accounting. It is the only method reaching non-DP accuracy on Abalone data and the first to reach it on Adults data, at a fraction of other methods’ privacy budget. At any point, the trade-off curve of our technique remains above others. Moreover, the test variance of our approach (computed over 10 trials) is very small because there is no noise added in the learning process.

Privacy loss bounds for the same levels of accuracy can be found in Table 1. The Bayesian accountant with its access to the distribution of gradients has a remarkable advantage. It is also worth mentioning, that for our methods we decreased $\delta$ to $10^{-5}$, because the failure probability $10^{-3}$ is too high for Adult dataset containing almost 50k samples.

7. Conclusion

In this paper, we consider the problem of differential privacy accounting in a generic machine learning context. We formalise the accounting workflow for a broad range of iterative learning algorithms and derive tight bounds on $\varepsilon, \delta$. Using this theory, we define two versions of a privacy accountant: generic and Bayesian. The latter produces better estimates under additional assumptions on data distribution.

Our evaluation shows that both the generic accountant and the Bayesian accountant produce tighter privacy loss estimates than the state-of-the-art moments accountant. For example, in deep learning context using the generic accountant could reduce the estimated failure probability $\delta$ from $10^{-5}$ to $10^{-10}$ for the same $\varepsilon$. Moreover, we demonstrate that a slight modification of variational inference leads to DP guarantees with almost no impact on accuracy. These guarantees become very strong when using data distribution for Bayesian accounting.
Improved Accounting for Differentially Private Learning

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.

Balle, B., Barthe, G., and Gaboardi, M. Privacy amplification by subsampling: Tight analyses via couplings and divergences. In Advances in Neural Information Processing Systems, pp. 6280–6290, 2018.

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

Bun, M., Dwork, C., Rothblum, G. N., and Steinke, T. Composable and versatile privacy via truncated cdps. In Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing, pp. 74–86. ACM, 2018.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. Journal of Machine Learning Research, 12(Mar):1069–1109, 2011.

Chen, L., Tao, C., Zhang, R., Henao, R., and Duke, L. C. Variational inference and model selection with generalized evidence bounds. In International Conference on Machine Learning, pp. 892–901, 2018.

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. Variational inference via χ upper bound minimization. In Advances in Neural Information Processing Systems, pp. 2732–2741, 2017.

Dimitrakakis, C., Nelson, B., Mitrokotsa, A., and Rubinstein, B. I. Robust and private bayesian inference. In International Conference on Algorithmic Learning Theory, pp. 291–305. Springer, 2014.

Dimitrakakis, C., Nelson, B., Zhang, Z., Mitrokotsa, A., and Rubinstein, B. I. Differential privacy for bayesian inference through posterior sampling. The Journal of Machine Learning Research, 18(1):343–381, 2017.

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.

Geumlek, J., Song, S., and Chaudhuri, K. Renyi differential privacy mechanisms for posterior sampling. In Advances in Neural Information Processing Systems, pp. 5289–5298, 2017.

Geyer, R. C., Klein, T., and Nabi, M. Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557, 2017.

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

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.

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

Kairouz, P., Oh, S., and Viswanath, P. Extremal mechanisms for local differential privacy. In Advances in neural information processing systems, pp. 2879–2887, 2014.

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

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.

Li, Y. and Turner, R. E. Rényi divergence variational inference. In Advances in Neural Information Processing Systems, pp. 1073–1081, 2016.

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.
Nissim, K., Raskhodnikova, S., and Smith, A. Smooth sensitivity and sampling in private data analysis. In Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, pp. 75–84. ACM, 2007.

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.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. Dp-em: Differentially private expectation maximization. arXiv preprint arXiv:1605.06995, 2016a.

Park, M., Foulds, J., Chaudhuri, K., and Welling, M. Variational bayes in private settings (vips). arXiv preprint arXiv:1611.00340, 2016b.

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.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In Global Conference on Signal and Information Processing (GlobalSIP), 2013 IEEE, pp. 245–248. IEEE, 2013.

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

Wang, Y.-X., Fienberg, S., and Smola, A. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In International Conference on Machine Learning, pp. 2493–2502, 2015.

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.