Quantifying identifiability to choose and audit ϵ in differentially private deep learning

Daniel Bernau
SAP SE
Karlsruhe, Germany
daniel.bernau@sap.com

Günther Eibl*
Salzburg University of Applied Sciences
Salzburg, Austria
guenther.eibl@en-trust.at

Philip W. Grassal*
University of Heidelberg
Heidelberg, Germany
philip-william.grassal@iwr.uni-heidelberg.de

Hannah Keller*
SAP SE
Karlsruhe, Germany
hannah.keller@sap.com

Florian Kerschbaum
University of Waterloo
Waterloo, Canada
florian.kerschbaum@uwaterloo.ca

ABSTRACT

Differential privacy allows bounding the influence that training data records have on a machine learning model. To use differential privacy in machine learning, data scientists must choose privacy parameters (ϵ, δ). Choosing meaningful privacy parameters is key, since models trained with weak privacy parameters might result in excessive privacy leakage, while strong privacy parameters might overly degrade model utility. However, privacy parameter values are difficult to choose for two main reasons. First, the theoretical upper bound on privacy loss (ϵ, δ) might be loose, depending on the chosen sensitivity and data distribution of practical datasets. Second, legal requirements and societal norms for anonymization often refer to individual identifiability, to which (ϵ, δ) are only indirectly related.

We transform (ϵ, δ) to a bound on the Bayesian posterior belief of the adversary assumed by differential privacy concerning the presence of any record in the training dataset. The bound holds for multidimensional queries under composition, and we show that it can be tight in practice. Furthermore, we derive an identifiability bound, which relates the adversary assumed in differential privacy to previous work on membership inference adversaries. We formulate an implementation of this differential privacy adversary that allows data scientists to audit model training and compute empirical identifiability scores and empirical (ϵ, δ).

PVLDB Reference Format:
Daniel Bernau, Günther Eibl, Philip W. Grassal, Hannah Keller, and Florian Kerschbaum. Quantifying identifiability to choose and audit ϵ in differentially private deep learning. PVLDB, 14(13): 3335-3347, 2021.
doi:10.14778/3484224.3484231

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/SAP-samples/security-research-identifiability-in-dpdl.

*Authors contributed equally to this work.

Proceedings of the VLDB Endowment, Vol. 14, No. 13 ISSN 2150-8097.
doi:10.14778/3484224.3484231

1 INTRODUCTION

The application of differential privacy (DP) for machine learning has received considerable attention by the privacy research community, leading to key contributions such as the tight estimation of privacy loss under composition [22, 23, 32] and differentially private stochastic gradient descent [1, 3, 41, 43] (DPSGD) for training neural networks. Still, data scientists must choose privacy parameters (ϵ, δ) to train a machine learning (ML) model using DPSGD. If the privacy parameters are stronger than necessary, model utility is sacrificed, as these parameters only formulate a theoretic upper bound that might not be reached when training an ML model with differentially private stochastic gradient descent on real-world data. If privacy parameters are too small, the trained model might be prone to reidentification attacks.

Several privacy regulations [37, 38] consider individual identifiability to gauge anonymization strength. Therefore, relatable scores that quantify reidentification risk to individuals can strongly affect the widespread implementation of anonymization techniques [34]. In consequence, if DP shall be used to comply with privacy regulations and find widespread adoption [36, 39], quantifying the resulting identifiability from privacy parameters (ϵ, δ) is required [7, 36].

Multiple approaches for choosing privacy parameters have been introduced, yet they do not reflect identifiability [2, 17], part from the original DP definition [5, 27, 40, 47], or lack applicability to common DP mechanisms for ML [26]. Especially in ML, practical membership inference (MI) attacks have been used to measure identifiability [5, 6, 16, 20, 21, 40, 42, 47]. However, MI adversaries are not assumed to have auxiliary information about the members of datasets that they aim to differentiate, which DP adversaries are assumed to possess. MI attacks thus offer intuition about the outcome of practical attacks; nonetheless, bounds on MI attacks in terms of ϵ are not tight [20], and consequently MI can only represent an empirical lower bound on identifiability.

Rather than analyzing the MI adversary, we consider a DP adversary with arbitrary auxiliary knowledge and derive maximum *Bayesian posterior belief* ρ_β as an identifiability bound related to (ϵ, δ), which bounds the adversary's certainty in identifying a member of the training data. Furthermore, we define the complementary score *expected membership advantage* ρ_α, which is related to the probability of success in a Bernoulli trial over the posterior beliefs.

ρ_α depends on the entire distribution of observed posterior beliefs, not solely the worst case posterior belief, and allows direct comparison with the membership advantage bound of Yeom et al. [47] for the MI adversary. We will show that the DP adversary achieves greater membership advantage than an MI adversary, implying that while both adversaries can be used to evaluate the protection of DP in machine learning, our implementable instance of the DP adversary comes closer to DP bounds.

A subsequent question is whether our identifiability bounds are tight in practice, since the factual guarantee (ϵ, δ) depends on the difference between possible input datasets [35]. In differentially private stochastic gradient descent, noise is scaled to global sensitivity, the maximum change that any single record in the training dataset is assumed to cause on the gradient during any training step. However, since all training data records are likely to be within the same domain (e.g., pictures of cars vs. pictures of nature scenes), global sensitivity might far exceed the difference between gradients over all training steps. We propose scaling the sensitivity to the difference between the gradients of a fixed dataset and any neighboring dataset and show for three reference datasets that we can indeed achieve tight bounds. Our main contributions are:

- Identifiability bounds for the posterior belief and expected membership advantage that are mathematical transformations of privacy parameters (ϵ, δ) and used in conjunction with RDP composition.
- The practical implementation of an adversary that meets all assumptions on worst-case adversaries against DP and allows us to audit DPSGD model instances w.r.t. to the empirical privacy loss besides enabling comparison with membership inference adversaries.
- A heuristic for scaling sensitivity in differentially private stochastic gradient descent. This heuristic leads to tight bounds on identifiability.

This paper is structured as follows. Preliminaries are presented in Section 2. We formulate identifiability scores and provide upper bounds on them in Sections 3 and 4. Section 5 specifies the application of these scores for a deep learning scenario, and we evaluate the scores for three deep learning reference datasets in Section 6. Section 7 discusses the practical relevance of our findings. We present related work and conclusions in Sections 8 and 9.

2 PRELIMINARIES

2.1 Differential Privacy

If the evaluation of a function $f : \mathcal{U} \rightarrow \mathcal{R}$ on a dataset $\mathcal{D}$ from domain $\mathcal{U}$ yields a result r, r inevitably leaks information about the entries $x \in \mathcal{D}$ (cf. impossibility of Dalenius' desideratum [9]). DP [9] offers an anonymization guarantee for statistical query functions $f(\cdot)$, perturbing r such that the result could have been produced from dataset $\mathcal{D}$ or some *neighboring* dataset $\mathcal{D}'$. A neighboring dataset $\mathcal{D}'$ either differs from $\mathcal{D}$ in the presence of one additional data point (unbounded DP) or in the value of one data point when a data point from $\mathcal{D}$ is replaced by another data point (bounded DP). In the context of this work, we will consider w.l.o.g. unbounded DP where $\mathcal{D}$ contains one data point x more than $\mathcal{D}'$ and $\mathcal{D} \setminus \mathcal{D}' = x$. To achieve differential privacy, noise is added to the result of $f(\cdot)$ by *mechanisms* $\mathcal{M}$ according to Definition 1. The impact of a single member $x \in \mathcal{D}$ on $f(\cdot)$ is bounded. If this impact is low compared to the noise specified by DP, plausible deniability is provided to this member of $\mathcal{D}$, even if $\mathcal{D}$ and the members' properties x (and thus also $\mathcal{D}'$) are known. For example, a single individual participating in a private analysis based on a census income dataset such as Adult [25] could therefore plausibly deny census participation and values of personal attributes. DP provides a strong guarantee, since it protects against a strong adversary with knowledge of up to all points in a dataset except one. As Definition 1 is an inequality, the privacy parameter ϵ can be interpreted as an upper bound on privacy loss.

Definition 1 ((ϵ, δ)-Differential Privacy [10]). *A mechanism $\mathcal{M}$ preserves (ϵ, δ)-differential privacy if for all independently sampled $\mathcal{D}, \mathcal{D}' \subseteq \mathcal{U}$, where $\mathcal{U}$ is a finite set, with $\mathcal{D}$ and $\mathcal{D}'$ differing in at most one element, and all possible mechanism outputs $\mathcal{S}$*

$$\Pr(\mathcal{M}(\mathcal{D}) \in \mathcal{S}) \leq e^{\epsilon} \cdot \Pr(\mathcal{M}(\mathcal{D}') \in \mathcal{S}) + \delta$$

The Gaussian mechanism is the predominant DP mechanism in ML for perturbing the outcome of stochastic gradient descent and adds noise independently sampled from a Gaussian distribution centered at zero. Prior work [11] has analyzed the tails of the normal distributions and found that bounding the standard deviation as follows fulfills (ϵ, δ)-DP:

$$\sigma > \Delta f \sqrt{2 \ln(1.25/\delta)}/\epsilon \tag{1}$$

Rearranged to solve for ϵ, this is:

$$\epsilon > \Delta f \sqrt{2 \ln(1.25/\delta)}/\sigma \tag{2}$$

σ depends not only on the DP guarantee, but also on a scaling factor Δf. Δf is commonly referred to as the sensitivity of a query function $f(\cdot)$ and comes in two forms: global sensitivity GS_f and local sensitivity LS_f. DP holds if mechanisms are scaled to GS_f of Definition 2, i.e., the maximum contribution of a record in the dataset to the outcome of $f(\cdot)$.

Definition 2 (Global Sensitivity). *Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring. For a given finite set $\mathcal{U}$ and function f the global sensitivity GS_f with respect to a distance function is*

$$GS_f = \max_{\mathcal{D}, \mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||$$

For the Gaussian mechanism, we use the global ℓ_2-sensitivity GS_{f_2}. Local sensitivity is specified in Definition 3 [35] and fixes dataset $\mathcal{D}$. Note that the absolute GS_f as of Definition 2 can also be defined relative to local sensitivity, as $GS_f = \max_{\mathcal{D}} LS_f(\mathcal{D})$. The impact of LS_f is that, compared to using GS_f, less noise is added when ϵ is held constant, and ϵ is decreased when the noise distribution is held constant.

Definition 3 (Local Sensitivity). *Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring. For a given finite set $\mathcal{U}$, independently sampled dataset $\mathcal{D} \subseteq \mathcal{U}$, and function f, the local sensitivity $LS_f(\mathcal{D})$ with respect to a distance function is*

$$LS_f(\mathcal{D}) = \max_{\mathcal{D}'} ||f(\mathcal{D}) - f(\mathcal{D}')||$$

In differentially private stochastic gradient descent, perturbed outputs are released repeatedly in an iterative process. $\mathcal{M}$ is represented by a differentially private version of an ML optimizer such

as Adam or SGD. The most basic form of accounting multiple data releases is sequential composition, which states that for a sequence of k mechanism executions each providing (ϵ_i, δ_i)-DP, the total privacy guarantee composes to $(\sum_i \epsilon_i, \sum_i \delta_i)$-DP; however, sequential composition adds more noise than necessary [1, 32].

A tighter analysis of composition is provided by Mironov [32]. (α, ϵ_{RDP})−Rényi differential privacy (RDP), with $\alpha > 1$ quantifies the difference in distributions $\mathcal{M}(\mathcal{D}), \mathcal{M}(\mathcal{D}')$ by their Rényi divergence [46]. For a sequence of k mechanism executions each providing $(\alpha, \epsilon_{RDP,i})$-RDP, the privacy guarantee composes to $(\alpha, \sum_i \epsilon_{RDP,i})$-RDP. The (α, ϵ_{RDP})-RDP guarantee converts to $(\epsilon_{RDP} - \frac{\ln \delta}{\alpha - 1}, \delta)$-DP. The Gaussian mechanism is calibrated to RDP by:

$$\epsilon_{RDP} = \alpha \cdot \Delta f^2 / 2\sigma^2 \tag{3}$$

2.2 Differentially Private Machine Learning

In machine learning a neural network (NN) is commonly provided a training dataset $\mathcal{D}$ where each of the data points $(x, y) \in \mathcal{D}$ consists of the features x and the label y. The goal is to learn a prediction function using an optimizer. A test set is used to evaluate generalization and utility of the trained model. This paper focuses on applying DP to stochastic gradient descent optimizers that output a gradient vector, which corresponds to the output of function f in DP. A variety of differentially private stochastic gradient descent (DPSGD) optimizers are available for deep learning, all of which depend on the privacy parameters (ϵ, δ) and the clipping norm $\mathtt{C}$ [1, 31]. DPSGD updates weights θ_i of the NN per training step $i \in k$ with $\theta_i \leftarrow \theta_{i-1} - \eta \cdot \tilde{g}_i$, where $\eta > 0$ is the learning rate. Differential privacy is achieved by perturbing the gradient $\tilde{g}_i = \mathcal{M}_i(\mathcal{D})$ with Gaussian noise. To limit the sensitivity Δf, the length of each per-example gradient is limited to the clipping norm $\mathtt{C}$ before perturbation, and the Gaussian perturbation is proportional to $\mathtt{C}$.

2.3 Membership Inference

Membership inference (MI) is a threat model for quantifying how accurately an adversary can identify members of the training data in ML. Yeom et al. [47] formalize MI in the following experiment:

Experiment 1. *(Membership Inference* $\mathrm{Exp}^{\mathrm{MI}}$*) Let* $\mathcal{A}_{\mathtt{MI}}$ *be an adversary,* $\mathcal{M}$ *be a differentially private learning algorithm,* n *be a positive integer, and* $\mathtt{Dist}$ *be a distribution over data points* (x, y)*. Sample* $\mathcal{D} \sim \mathtt{Dist}^n$ *and let* $\mathbf{r} = \mathcal{M}(\mathcal{D})$*. The membership experiment proceeds as follows:*

(1) Sample $z_{\mathcal{D}}$ *uniformly from* $\mathcal{D}$ *and* $z_{\mathtt{Dist}}$ *from* $\mathtt{Dist}$
(2) Choose $b \leftarrow \{0, 1\}$ *uniformly at random*
(3) Let

$$z = \begin{cases} z_{\mathcal{D}} & \text{if } b{=}1 \\ z_{\mathtt{Dist}} & \text{if } b{=}0 \end{cases}$$

(4) $\mathcal{A}_{\mathtt{MI}}$ *outputs* $b' = \mathcal{A}_{\mathtt{MI}}(\mathbf{r}, z, \mathtt{Dist}, n, \mathcal{M}) \in \{0, 1\}$*. If* $b' = b$*,* $\mathcal{A}_{\mathtt{MI}}$ *succeeds and the output of the experiment is 1. It is 0 otherwise.*

2.4 Differential Identifiability and the relation to the DP adversary

Lee et al. [26, 27] introduce differential identifiability (DI) as a strong inference threat model. DI assumes that the adversary calculates the likelihood of all possible input datasets, so-called *possible worlds* in a set Ψ, given a mechanism output r. Li et al. [28] show that the DI threat model maps to the worst case against which bounded DP protects when $|\Psi| = 2$, since DP considers two neighboring datasets $\mathcal{D}, \mathcal{D}'$ by definition. The DI experiment $\mathrm{Exp}^{\mathrm{DI}}$ is similar to $\mathrm{Exp}^{\mathrm{MI}}$, since the adversary must decide whether the dataset contains the member that differs between the known $\mathcal{D}'$ and $\mathcal{D}$, or not. For comparison we reformulate DI as a cryptographic experiment:

Experiment 2. *(Differential Identifiability* $\mathrm{Exp}^{\mathrm{DI}}$*) Let* $\mathcal{A}_{\mathtt{DI}}$ *be an adversary,* $\mathcal{M}$ *be a differentially private learning algorithm,* $\mathcal{D}$ *and* $\mathcal{D}'$ *be neighboring datasets drawn mutually independently from distribution* $\mathtt{Dist}$*, using either bounded or unbounded definitions. The differential identifiability experiment* $\mathrm{Exp}^{\mathrm{DI}}$ *proceeds as follows:*

(1) Set $\mathbf{r}_D := \mathcal{M}(\mathcal{D})$ *and* $\mathbf{r}_{\mathcal{D}'} := \mathcal{M}(\mathcal{D}')$
(2) Choose $b \leftarrow \{0, 1\}$ *uniformly at random*
(3) Let

$$\mathbf{r} = \begin{cases} \mathbf{r}_{\mathcal{D}}, & \text{if } b = 1 \\ \mathbf{r}_{\mathcal{D}'}, & \text{if } b = 0 \end{cases}$$

(4) $\mathcal{A}_{\mathtt{DI}}$ *outputs* $b' = \mathcal{A}_{\mathtt{DI}}(\mathbf{r}, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist}) \in \{0, 1\}$*. If* $b' = b$*,* $\mathcal{A}_{\mathtt{DI}}$ *succeeds and the output of the experiment is 1. It is 0 otherwise.*

Since Experiment 2 precisely defines an adversary with access to arbitrary background knowledge of up to all but one record in $\mathcal{D}$ and $\mathcal{D}'$, $\mathcal{A}_{\mathtt{DI}}$ is an implementable instance of the DP adversary [12]. Compared to the MI adversary, the DI adversary is stronger, since $\mathcal{A}_{\mathtt{DI}}$ knows the alternative dataset $\mathcal{D}'$ instead of only the distribution $\mathtt{Dist}$ from which $\mathcal{D}'$ was chosen. The experiment defined above is general and applies to deep learning using gradient descent as follows: the knowledge of the mechanism $\mathcal{M}$ implies knowledge about the architecture of the NN and the learning parameters $\eta, \mathtt{C}$, as well as number of iterations k. The experiment is formulated s.t. it could be applied for a single iteration, and the output $\mathbf{r}$ of the mechanism is the perturbed gradient $\tilde{g}_i$ from iteration i of the NN training. However, after the entire learning process, consisting of k rounds, $\mathcal{A}_{\mathtt{DI}}$ has more information $R_k = (\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k)$ and therefore a higher chance to win Experiment 2. In this case, the same value of b is chosen in every round, since the training data is kept constant over all learning steps. This is the standard case considered in our paper and motivates the need for composition theorems. According Experiment 2, the DI adversary could know almost all of the training data from a public dataset of census data, for example, and observe the NN gradient updates at every training step. The assumption that $\mathcal{A}_{\mathtt{DI}}$ has access to all gradients during learning may seem overly strong; however, this setting is of theoretical interest, since the bounds that we prove for the DI adversary $\mathcal{A}_{\mathtt{DI}}$ will also hold for weaker adversaries. Furthermore, the assumptions can be fulfilled in federated learning, for example. In federated learning, multiple data owners jointly train a global model by sharing gradients for their individual training data subsets with a central aggregator. The aggregator combines the gradients and shares the

aggregated update with all data owners. If $\mathcal{A}_{\mathtt{DI}}$ participates as a data owner, $\mathcal{A}_{\mathtt{DI}}$ is able to observe joint model updates.

3 IDENTIFIABILITY SCORES FOR DP

In this section we formulate two scores for identifiability of individual training records when releasing a differentially private NN. The scores are compatible with DP under multidimensional queries and composition. However, we first show that if Adv^{DI} is bounded, then this bound also holds for Adv^{MI}. Equivalently, $\mathcal{A}_{\mathtt{DI}}$ is stronger than $\mathcal{A}_{\mathtt{MI}}$ due to additional available auxiliary information. Concretely, $\mathcal{A}_{\mathtt{DI}}$ knows both neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ instead of only receiving one value z and the size n of the dataset from which the data points are drawn. This leads us to Proposition 1, which we formally prove by reduction in an extended version of this paper [4].

Proposition 1. *DI implies MI: if $\mathcal{A}_{\mathtt{MI}}$ wins Exp^{MI}, then one can construct $\mathcal{A}_{\mathtt{DI}}$ that wins Exp^{DI}.*

We define *posterior belief* β, which quantifies identifiability for iterative mechanisms in Section 3.1. Second, we define *membership advantage* Adv^{DI} for $\mathcal{A}_{\mathtt{DI}}$ in Section 3.2, which is a complementary identifiability score offering a scaled quantification of the adversary's probability of success.

3.1 Posterior Belief in Identifying the Training Dataset

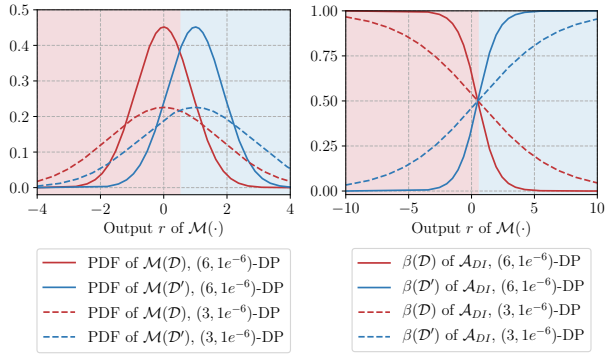
To quantify individual identifiability from privacy parameters (ϵ, δ), we use the Bayesian posterior belief. After having observed gradients R_k, the adversary $\mathcal{A}_{\mathtt{DI}}$ can update the probabilities for both the training dataset $\mathcal{D}$ and the alternate dataset $\mathcal{D}'$, that differs from $\mathcal{D}$ in an individual record $x = \mathcal{D} \setminus \mathcal{D}'$. The posterior belief quantifies the certainty with which $\mathcal{A}_{\mathtt{DI}}$ is able to identify the training dataset used by a NN and consequently the presence of the individual record x. This belief is formulated as a conditional probability depending on observations R_k during training. For a census dataset such as Adult, the posterior belief measures the probability that a particular individual x participated in the census after observing training using data $\mathcal{D}$. Since this belief has an upper bound for each possible member x of the dataset, no member of $\mathcal{D}$ can be identified. Posterior belief therefore relates theoretical DP privacy guarantees to privacy regulations and societal norms through its identifiability formulation, since the noise, and therefore the posterior belief, depends on (ϵ, δ).

Definition 4 (Posterior Belief). *Consider the setting of Experiment 2 and denote $R_k = (\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k)$ as the result matrix, comprising k multidimensional mechanism results. The posterior belief in the correct dataset $\mathcal{D}$ is defined as the probability conditioned on all the information observed during the adaptive computations*

$$\beta_k := \Pr(\mathcal{D}|R_k) = \frac{\Pr(\mathcal{D}, R_k)}{\Pr(\mathcal{D}, R_k) + \Pr(\mathcal{D}', R_k)}$$

where the probability $\Pr(\mathcal{D}|R_k)$ is over the random iterative choices of the mechanisms up to step k.

Each β_k can be computed from the previous β_{k-1}. The final belief can be computed using Lemma 1, which we will use to further analyze the strongest possible attacker $\mathcal{A}_{\mathtt{DI}}$ of Experiment 2. The proof for Lemma 1 is available in the extended version of this paper [4].



(a) PDFs **(b) Posterior belief**

Figure 1: The decision boundary of $\mathcal{A}_{\mathtt{DI}}$

Lemma 1 (Calculation of the posterior belief). *Assuming uniform priors and independent mechanisms $\mathcal{M}_i$ (more precisely, the noise of the mechanisms must be sampled independently), the posterior belief on dataset $\mathcal{D}$ can be computed as*

$$\begin{aligned}\beta_k &= \frac{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}) = \mathbf{r}_i)}{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}) = \mathbf{r}_i) + \prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}') = \mathbf{r}_i)} \\ &= \frac{1}{1 + \frac{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}') = \mathbf{r}_i)}{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}) = \mathbf{r}_i)}}\end{aligned}$$

In our analysis $\mathcal{A}_{\mathtt{DI}}$ is a binary classifier that chooses the label with the highest posterior probability β_k. If prior beliefs are uniform, this decision process can be simplified. Consider $X_1 := \mathcal{M}(\mathcal{D})$ and $X_0 := \mathcal{M}(\mathcal{D}')$. Since $\mathcal{A}_{\mathtt{DI}}$ knows $\mathcal{D}, \mathcal{D}'$ and $\mathcal{M}$, $\mathcal{A}_{\mathtt{DI}}$ also knows the corresponding probability densities g_{X_1} and g_{X_0}. The densities are identical and defined by $\mathcal{M}$, but are centered at the different results $f(\mathcal{D})$ and $f(\mathcal{D}')$, respectively, as visualized in Figure 1a with $f(\mathcal{D}) = 0, f(\mathcal{D}') = 1$. When $\mathcal{A}_{\mathtt{DI}}$ has equal prior beliefs, $\mathcal{A}_{\mathtt{DI}}$ decides whether R_k is more likely to stem from X_1 or X_0 and therefore chooses

$$\begin{aligned}\mathcal{A}_{\mathtt{DI}}(R_k, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist}) &= \underset{D \in \{\mathcal{D}, \mathcal{D}'\}}{\arg\max}\ \beta(D|R_k) \\ &= \underset{b \in \{0,1\}}{\arg\max}\, g_{X_b}(R_k)\end{aligned} \tag{4}$$

$\beta(\mathcal{D})$ and $\beta(\mathcal{D}')$ for our example are visualized in Figure 1b. $\mathcal{A}_{\mathtt{DI}}$ acts as a naive Bayes classifier whose decision is depicted by the background color. The input features are the perturbed results R_k, and the exact probability distribution of each class is known. The distributions are entirely defined by $\mathcal{D}$, $\mathcal{D}'$, and $\mathcal{M}$, so $\mathcal{A}_{\mathtt{DI}}$ does not use the knowledge of `Dist`. The posterior belief quantifies the probability of R_k; however, in another instance, R_k could differ. In Section 4.1, we will therefore define an upper bound on $\beta(\mathcal{D})$.

3.2 Advantage in Identifying the Training Dataset

The posterior belief β_k quantifies the probability of inferring membership of a single record x. For example, when β_k is low for a census dataset, the individual x can plausibly deny presence in $\mathcal{D}$, and thus presence in the census. In practice, it is also important to know how often $\mathcal{A}_{\texttt{DI}}$ makes a correct guess, which only occurs when $\beta_k > 0.5$. This is quantified by the advantage, which is the success rate normalized to the range $[-1, 1]$, where $Adv = 0$ corresponds to random guessing. Membership advantage was introduced to quantify the success of $\mathcal{A}_{\texttt{MI}}$ [47]; however, its definition can be used for $\mathcal{A}_{\texttt{DI}}$ of Exp^{DI}. Generically:

Definition 5 (Advantage). *Given an experiment Exp the advantage is defined as*

$$Adv = 2\Pr(Exp = 1) - 1$$

where the probability is over the random iterative choices of the mechanisms up to step k. The advantage in Exp^{DI} *is denoted* Adv^{DI}*, while the advantage in* Exp^{MI} *is* Adv^{MI}.

4 DERIVATION OF UPPER BOUNDS

Within this section we use the DP guarantee to derive upper bounds for *posterior belief* and *advantage* in Sections 4.1 and 4.2. In Section 4.3, we define expected membership advantage for the Gaussian mechanism, since the original bound is loose.

4.1 Upper Bound for the Posterior Belief

We formulate a generic bound on the Bayesian posterior belief that is independent of datasets $\mathcal{D}$ and $\mathcal{D}'$, the mechanism $\mathcal{M}$, and the result matrix $R = (\mathbf{r}_0, \mathbf{r}_1, \ldots, \mathbf{r}_k)$ comprising k multidimensional mechanism outputs. The proposed bound solely assumes that the DP bound holds and makes no further simplifications, which results in an identifiability-based interpretation of DP guarantees. Theorem 2 shows that $\mathcal{A}_{\texttt{DI}}$ operates under the sequential composition theorem, for both for ϵ-DP and for (ϵ, δ)-DP. We refer the interested reader to an extended version of this work for the proof [4].

Theorem 2 (Bounds for the Adaptive Posterior Belief). *Consider experiment* Exp^{DI} *with neighboring datasets* $\mathcal{D}$ *and* $\mathcal{D}'$*. Let* $\mathcal{M}_1, \ldots, \mathcal{M}_k$ *be a sequence of arbitrary but independent differentially private learning algorithms.*
(i) Each $\mathcal{M}_i$ *provides* $\epsilon_1, \ldots, \epsilon_k$*-DP to functions* f_i *with multidimensional output. Then the posterior belief of* $\mathcal{A}_{\texttt{DI}}$ *is bounded by*

$$\beta_k(\mathcal{D}|R_k) \leq \rho_\beta = \frac{1}{1 + e^{-\sum_{i=1}^{k} \epsilon_i}}$$

(ii) Each $\mathcal{M}_i$ *provides* (ϵ_i, δ_i)*-DP to multidimensional functions* f_i*. Then the same bound as above holds with probability* $1 - \sum_{i=1}^{k} \delta_i$.

Equivalently one can specify a desired posterior belief and calculate the overall ϵ, which can be spent on a composition of differentially private queries:

$$\epsilon = \ln\left(\frac{\rho_\beta}{1 - \rho_\beta}\right) \tag{5}$$

The value for δ can be chosen independently according to the recommendation that $\delta \ll \frac{1}{N}$ with N points in the input dataset [11].

4.2 Upper Bound for the Advantage in Identifying the Training Dataset for General Mechanisms

We now formulate an upper bound for the advantage Adv^{DI} of $\mathcal{A}_{\texttt{DI}}$ in Proposition 2. The membership advantage of $\mathcal{A}_{\texttt{MI}}$ has been bounded in terms of ϵ and defines $\mathcal{A}_{\texttt{MI}}$'s success [47]. The general bound for $\mathcal{A}_{\texttt{MI}}$ also holds for $\mathcal{A}_{\texttt{DI}}$ based on Proposition 1. Again, a proof is provided in the extended version [4].

Proposition 2 (Bound on the Expected Membership Advantage for $\mathcal{A}_{\texttt{DI}}$). *For any* ϵ*-DP mechanism the identification advantage of* $\mathcal{A}_{\texttt{DI}}$ *in experiment* Exp^{DI} *can be bounded as*

$$\text{Adv}^{\text{DI}} \leq (e^{\epsilon} - 1)\Pr(\mathcal{A}_{\texttt{DI}} = 1|b = 0)$$

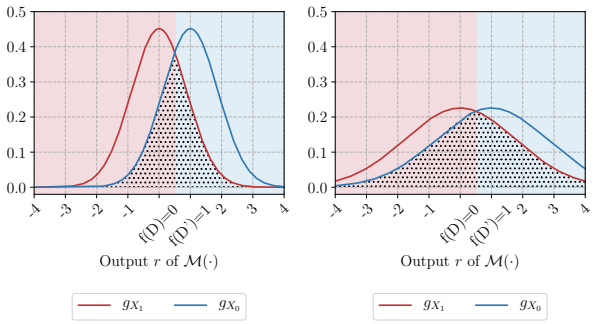
Bounding $\Pr(\mathcal{A}_{\texttt{DI}} = 1|b = 0)$ by 1 results in $\text{Adv}^{\text{DI}} \leq e^{\epsilon} - 1$. When $\mathcal{A}_{\texttt{DI}}$ acts as a naive Bayes classifier, only a complete lack of utility from infinite noise results in $\Pr(\mathcal{A}_{\texttt{DI}} = 1|b = 0) = 0.5$. Otherwise, $\Pr(\mathcal{A}_{\texttt{DI}} = 1|b = 0) \ll 0.5$; therefore, the membership advantage bound is usually not tight. This is in line with Jayaraman et al. [20] who expect that this would be the case for MI.

4.3 Upper Bound for the Advantage in Identifying the Training Dataset for Gaussian Mechanisms

In practice, $\mathcal{A}_{\texttt{DI}}$ will be faced with a specific DP mechanism, and we focus on the mechanism used in DPSGD to find a tighter bound than the generic bound described in the previous section. We use the notation $\mathcal{A}_{\texttt{DI,Gau}}$ and $\text{Adv}^{\text{DI,Gau}}$ to specify the adversary and advantage of an instantiation of $\mathcal{A}_{\texttt{DI}}$ against the Gaussian mechanism with (ϵ, δ)-DP. We now derive a tighter bound ρ_α on $\text{Adv}^{\text{DI,Gau}}$. Note that under the assumption of equal priors, the strongest possible adversary of Eq. (4) chooses $b = 1$ if $(g_{X_1}(\mathbf{r}) - g_{X_0}(\mathbf{r})) > 0$ and $b = 0$ otherwise. The resulting bound on $\text{Adv}^{\text{DI,Gau}}$ is constructed from $\mathcal{A}_{\texttt{DI,Gau}}$'s strategy; however, the bound holds for all weaker adversaries, including $\mathcal{A}_{\texttt{MI}}$. Since we argue that $\mathcal{A}_{\texttt{DI,Gau}}$ precisely represents the assumptions of DP, the bound should hold for other possible attacks in the realm of DP and the Gaussian mechanism under the i.i.d. assumption.

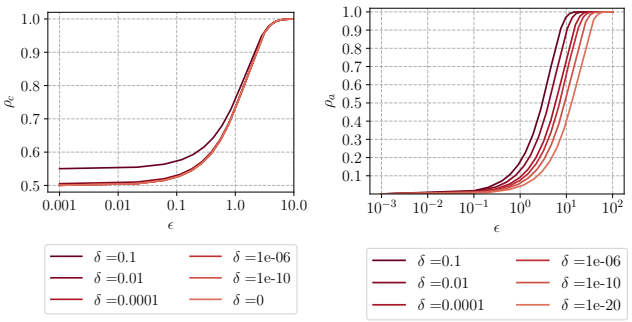
Since $\mathcal{A}_{\texttt{DI,Gau}}$ is a naive Bayes classifier with known probability distributions, we use the properties of normal distributions (we refer to Tumer et al. [45] for full details). We find that the decision boundary does not change under the Gaussian mechanism $\mathcal{M}_{Gau}$ with different (ϵ, δ) guarantees as long as the probability density functions (PDF) are symmetric. Holding $\mathcal{M}(\mathcal{D}) = r$ constant and reducing (ϵ, δ) solely affects the posterior belief of $\mathcal{A}_{\texttt{DI,Gau}}$, not the choice of $\mathcal{D}$ or $\mathcal{D}'$. For example, consider the example of Figure 2. If a $(6, 10^{-6})$-DP $\mathcal{M}_{Gau}$ is applied for perturbation, $\mathcal{A}_{\texttt{DI,Gau}}$ has to choose between the two PDFs in Figure 2a. Increasing the privacy guarantee to $(3, 10^{-6})$-DP in Figure 2b squeezes the PDFs and belief curves. The corresponding regions of error are shaded in Figures 2a and 2b, where we see that a stronger guarantee reduces $\text{Adv}^{\text{DI,Gau}}$.

We assume throughout this paper that $\mathcal{A}_{\texttt{DI,Gau}}$ has uniform prior beliefs on the possible databases $\mathcal{D}$ and $\mathcal{D}'$. This distribution is iteratively updated based on the posterior resulting from the mechanism output r. If $\mathcal{M}_{Gau}$ is used to achieve (ϵ, δ)-DP, we can



(a) Error regions $(6, 1e^{-6})$-DP **(b) Error regions $(3, 1e^{-6})$-DP**

Figure 2: Error regions for varying ϵ, $\mathcal{M}_{Gau}$



(a) ρ_β **(b) ρ_α**

Figure 3: ρ_β and ρ_α for various (ϵ, δ) when using $\mathcal{M}_{Gau}$

determine the expected membership advantage of the practical attacker $\mathcal{A}_{\text{DI,Gau}}$ analytically by the overlap of the resulting Gaussian distributions [29, p. 321]. We thus consider two multidimensional Gaussian PDFs (i.e., $\mathcal{M}(\mathcal{D})$, $\mathcal{M}(\mathcal{D}')$) with covariance matrix Σ and means (without noise) $\mu_1 = f(\mathcal{D}), \mu_2 = f(\mathcal{D}')$. This leads us to Theorem 3. We again refer to the extended version of this paper [4] for the proof.

Theorem 3 (Tight Bound on the Expected Adversarial Membership Advantage). *For the (ϵ, δ)-differentially private Gaussian mechanism, the expected membership advantage of the strong probabilistic adversary on either dataset $\mathcal{D}, \mathcal{D}'$.*

$$\text{Adv}^{\text{DI}} \leq \rho_\alpha = 2\Phi\left(\frac{\epsilon}{2\sqrt{2\ln(1.25/\delta)}}\right) - 1$$

where Φ is the cumulative density function of the standard normal distribution.

We can calculate ϵ from a chosen maximum expected advantage

$$\epsilon = \sqrt{2\ln(1.25/\delta)}\,\Phi^{-1}\left(\frac{\rho_\alpha + 1}{2}\right) \tag{6}$$

(ϵ, δ) guarantees with $\delta > 0$ can be expressed via a scalar value ρ_α. Summarizing, we now have complementary interpretability scores, where ρ_β represents a bound on individual deniability and ρ_α relates to the expected probability of reidentification. While ρ_β holds for all mechanisms, ρ_α was derived solely for the Gaussian mechanism. We provide example plots of ρ_β and ρ_α for different (ϵ, δ) in Figure 3. To compute both scores, we use Theorems 2 and 3. We set $f(\mathcal{D}) = (0_1, 0_2, \ldots, 0_k)$ and $f(\mathcal{D}') = (1_1, 1_2, \ldots, 1_k)$ for all dimensions k, so $GS_{f_2} = \sqrt{k}$. Figure 3a illustrates that there is no significant difference for ρ_β between ϵ-DP and (ϵ, δ)-DP. In contrast, ρ_α strongly depends on the choice of δ.

4.4 RDP Instead of Sequential Composition

In iterative settings, such as NN training, the data scientist will have to perform multiple mechanism executions, which necessitates the use of composition theorems to split the total guarantee into guarantees per iteration (ϵ_i, δ_i). Sequential composition only offers loose bounds in practice [13, 23]; we suggest using RDP composition, which allows a tight analysis of the privacy loss over a series of mechanisms. Therefore, we adapt both ρ_β and ρ_α to RDP.

We first demonstrate that RDP composition results in stronger (ϵ, δ) guarantees than sequential composition for a fixed bound ρ_β. We start from the simple RDP belief:

$$\begin{aligned}
\beta_k(\mathcal{D}|R) &\leq \frac{1}{1 + \prod_{i=1}^{k} e^{-(\epsilon_{RDP,i} + (\alpha-1)^{-1}\ln(1/\delta_i))}} \\
&= \frac{1}{1 + e^{k(\alpha-1)^{-1}\ln(\delta_i) - \sum_{i=1}^{k}\epsilon_{RDP,i}}} \qquad (7) \\
&= \frac{1}{1 + e^{(\alpha-1)^{-1}\ln(\delta_i^k) - \sum_{i=1}^{k}\epsilon_{RDP,i}}} \\
&= \frac{1}{1 + e^{-(\sum_{i=1}^{k}\epsilon_{RDP,i} - (\alpha-1)^{-1}\ln(\delta_i^k))}} = \rho_\beta \qquad (8)
\end{aligned}$$

We assume the same value of δ_i is used during every execution and can therefore remove it from the sum in Eq. (7). Eq. (8) and the conversion (α, ϵ_{RDP})-RDP to $(\epsilon_{RDP} - \frac{\ln\delta}{\alpha-1}, \delta)$-DP imply that an RDP-composed bound can be achieved with a composed δ equal to δ_i^k. We know that sequential composition results in a composed δ value equal to $k\delta_i$. Since $\delta^k < k\delta$, RDP offers a stronger (ϵ, δ) guarantee for the same ρ_β, and results in a tighter bound for ρ_β under composition. This behavior can also be interpreted as the fact that holding the composed (ϵ, δ) guarantee constant, the value of ρ_β is greater when sequential composition is used compared to RDP.

A similar analysis of the expected membership advantage under composition is required when considering a series of mechanisms $\mathcal{M}$. We restrict our elucidations to the Gaussian mechanism. The k-fold composition of $\mathcal{M}_{Gau_i}$, each step guaranteeing $(\alpha, \epsilon_{RDP,i})$-RDP, can be represented by a single execution of $\mathcal{M}_{Gau}$ with k-dimensional output guaranteeing $(\alpha, \epsilon_{RDP} = k\epsilon_{RDP,i})$-RDP. We use Eq. (3) and the fact that GS_{f_2} bounds $\|\mu_{1,i} - \mu_{2,i}\|$.

$$\begin{aligned}
\text{Adv}^{\text{DI,Gau}} &= 2\Phi\left(\frac{\|\mu_1 - \mu_2\|_2}{2\sigma_i}\right) - 1 = 2\Phi\left(\frac{\sqrt{k}\|\mu_{1,i} - \mu_{2,i}\|_2}{2GS_{f_2}\sqrt{\alpha/(2\epsilon_{RDP,i})}}\right) - 1 \\
&\leq 2\Phi\left(\frac{\sqrt{k}}{2\sqrt{\alpha/(2\epsilon_{RDP,i})}}\right) - 1 = 2\Phi\left(\sqrt{\frac{k\epsilon_{RDP,i}}{2\alpha}}\right) - 1 \\
&= 2\Phi\left(\sqrt{\frac{\epsilon_{RDP}}{2\alpha}}\right) - 1 = \rho_\alpha
\end{aligned}$$

The result shows that $\mathcal{A}_{\texttt{DI,Gau}}$ fully takes advantage of the RDP composition properties of $\epsilon_{RDP,i}$ and α; as expected, ρ_α takes on the same value, regardless of whether k composition steps with $\epsilon_{RDP,i}$ or a single composition step with ϵ_{RDP} is carried out. Therefore, we can calculate the final ρ_α for functions with multiple iterations, such as the training of deep learning models, and ρ_α can be decomposed into a privacy guarantee per composition step with RDP.

5 APPLICATION TO DEEP LEARNING

In DPSGD, the stochastic gradient descent optimizer adds Gaussian noise with standard deviation σ to the computed gradients. The added noise ensures that the learned NN is (ϵ, δ) differentially private w.r.t. the training dataset. This section illustrates our method for choosing DPSGD privacy parameters. Data scientists may first choose upper bounds for the posterior belief, from which ϵ is obtained using Eq. (5). From ϵ and the sensitivity, the standard deviation σ of the Gaussian noise is determined.

We discuss a heuristic for estimating the local sensitivity in Section 5.1. Then, Section 5.2 formulates an algorithm for implementing $\mathcal{A}_{\texttt{DI,Gau}}$, and discusses how this algorithm is used to empirically quantify the posterior belief and the advantage. Finally, using the implemented adversary $\mathcal{A}_{\texttt{DI,Gau}}$ a method for auditing the privacy loss ϵ and the bounds derived in Section 4 is provided in Section 5.3.

5.1 Setting Privacy Parameters and Determining the Sensitivity

Based on the recommendation to set C to the median of the norms of unclipped gradients [1] we set $\texttt{C} = 3$ in all experiments. In the following, we describe how to determine the standard deviation of Gaussian noise σ. We want to limit $\mathcal{A}_{\texttt{DI,Gau}}$'s belief of distinguishing a training dataset differing in any chosen person by setting the upper bound for the posterior belief ρ_β. We then transform ρ_β to an overall ϵ for the k update steps in DPSGD using Eq. (5), which in turn leads to σ for the DPSGD using Eq. (1). In Eq. (1) two parameters need to be set: Δf and δ. While we set δ to $1/|\mathcal{D}|$ for all experiments, the choice of Δf is more challenging. The upper bound for the privacy loss ϵ can only be reached when Δf is set specifically to the sensitivity of the dataset at hand. We calculate the local sensitivity for bounded DP as

$$LS_{\hat{g}_i}(\mathcal{D}) = n \cdot ||\hat{g}_i(\mathcal{D}') - \hat{g}_i(\mathcal{D})||,$$

and for unbounded DP as

$$LS_{\hat{g}_i}(\mathcal{D}) = ||(n-1) \cdot \hat{g}_i(\mathcal{D}') - n \cdot \hat{g}_i(\mathcal{D})||,$$

where $\hat{g}_i(\mathcal{D})$ and $\hat{g}_i(\mathcal{D}')$ represent the average of all clipped, unperturbed per-example gradients $\bar{g}_i(x) \forall x \in \mathcal{D}$ and $x \in \mathcal{D}'$.

Since clipping is done before perturbation, the global sensitivity GS_f in DPSGD is set to the clipping norm for unbounded DP, i.e., $GS_f = \texttt{C}$. The sensitivity bounds the impact of a data point on the total gradient, equivalent to the difference between the gradients differing between $\mathcal{D}$ and $\mathcal{D}'$, which is artificially bounded by C for unbounded DP. For bounded DP where one record is instead replaced with another in $\mathcal{D}'$, the lengths of the clipped gradients of these two records could each be C and point in opposite directions resulting in $n \cdot ||\hat{g}_i(\mathcal{D}') - \hat{g}_i(\mathcal{D})||_2 \leq 2\texttt{C}$.

Although C bounds the influence of a single training record on the gradient, C may well be loose, since C does not necessarily reflect the factual difference between the training dataset and possible neighboring datasets. When C is loose, the DP bound on privacy loss ϵ is not reached, and the identifiability metrics ρ_α and ρ_β will not be reached either. Nissim et al. [35] proposed local sensitivity LS_f to specifically scale noise to the input data. The use of LS_f decreases the noise scale by narrowing the DP guarantee from protection against inference on any possible adjacent datasets to inference on the original dataset and any adjacent dataset. In ML projects training and test data are often sampled from a static holdout, where all data points stem from a domain of similar data. If the holdout is a very large dataset, only the specific neighboring datasets possible in this domain need to be protected under DP. To reach the DP bound, we suggest fixation of the training dataset $\mathcal{D}$ and considering only neighboring datasets $\mathcal{D}'$ adjacent to $\mathcal{D}$.

However, approximating $LS_{\hat{g}_i}$ for NN training is difficult because the gradient function output depends not only on $\mathcal{D}$ and $\mathcal{D}'$, but also on the architecture and current weights of the network. To ease this dilemma, we propose *dataset sensitivity* in Definition 6. Dataset sensitivity is a heuristic with which we strive to consider the neighboring dataset $\hat{\mathcal{D}}'$ with the largest difference to $\mathcal{D}$ within the overall ML dataset $\mathcal{U}$ in an effort to approximate $LS_{\hat{g}_i}$. We assume that similar data points will result in similar gradients. While this assumption does not necessarily hold under crafted adversarial examples [15], for which privacy protection cannot be guaranteed, the malicious intent renders the necessity for their protection debatable. In Definition 6 the dissimilarity measure of specific datasets is not further specified.

Definition 6 (Dataset Sensitivity). *Consider a given dataset $\mathcal{U}$, a training dataset $\mathcal{D} \subseteq \mathcal{U}$, all neighboring datasets $\mathcal{D}' \subseteq \mathcal{U}$ and a dissimilarity measure d. The dataset sensitivity $DS(\mathcal{D})$ w.r.t. dissimilarity measure d is then defined as*

$$DS(\mathcal{D}) = \max_{\mathcal{D}'} d(\mathcal{D}, \mathcal{D}')$$

and consequently

$$\hat{\mathcal{D}}' := \arg\max_{\mathcal{D}'} d(\mathcal{D}, \mathcal{D}')$$

In practice, if a dissimilarity or distance measure d of individual data points is available, it can be used to find the most dissimilar neighboring dataset $\hat{\mathcal{D}}'$ that maximizes the dataset sensitivity. The computation of $\mathcal{D}'$ depends on the neighboring datasets and is different for unbounded and bounded DP. More precisely, for unbounded DP one forms $\hat{\mathcal{D}}' = \mathcal{D} \setminus \{x'\}$ by removing the most dissimilar data point $\hat{x}$ from the training data

$$\hat{x} = \arg\max_{x_1 \in \mathcal{D}} \sum_{x_2 \in \mathcal{D} \setminus x_1} d(x_1, x_2) \tag{9}$$

The dataset $\hat{\mathcal{D}}'$ is then used to approximate the local sensitivity $LS_{\hat{g}_i}$ by

$$LS_{\hat{g}_i}(\mathcal{D}) \approx \hat{LS}_{\hat{g}_i}(\mathcal{D}) := \|\bar{g}_i(\hat{x})\|, \tag{10}$$

where $\bar{g}_i(x)$ is the clipped gradient of data point x in step i. The simplification from $LS_{\hat{g}_i}$ to DS allows us to bypass the complex gradient calculations to identify dissimilar $\mathcal{D}$ and $\mathcal{D}'$. The computational complexity of computing the dataset sensitivity only depends on the dataset size n, but not the number of iterations k, like the local sensitivity does. For bounded DP where a neighboring

dataset is formed by replacing an element $\{x\} \in \mathcal{D}$ with an element $x' \in \mathcal{U} \setminus \mathcal{D}$ one searches for

$$(\hat{x}, \hat{x}') = \underset{x \in \mathcal{D}, x' \in \mathcal{U} \setminus \mathcal{D}}{\arg\max} d(x, x'). \tag{11}$$

and approximates the local sensitivity as

$$LS_{\hat{g}_i}(\mathcal{D}) \approx \hat{LS}_{\hat{g}_i}(\mathcal{D}) := \|\bar{g}_i(\hat{x}) - \bar{g}_i(\hat{x}')\| \tag{12}$$

5.2 Empirical Quantification of Posterior Beliefs and Advantages

In Section 5.1 the noise scale σ limits the upper bound for the posterior belief of $\mathcal{A}_{\mathtt{DI}}$ on the original dataset $\mathcal{D}$. According to Theorem 2 this upper bound holds with probability $1 - \delta$. For a given dataset, the posterior belief might be much smaller than the bound, so it is desirable to determine the empirical posterior belief on $\mathcal{D}$. The same holds for the advantage Adv^{DI} and the upper bound ρ_α from Theorem 3 w.r.t. identifying dataset $\mathcal{D}$. We formulate an implementation of the adversary $\mathcal{A}_{\mathtt{DI,Gau}}$ which allows us to assess the empirical posterior belief β and membership advantage Adv^{DI}, and thus the empirical privacy loss of specific trained models.

The adversary $\mathcal{A}_{\mathtt{DI,Gau}}$ strives to identify the training dataset, having the choice between neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$. In addition to $\mathcal{D}$ and $\mathcal{D}'$, $\mathcal{A}_{\mathtt{DI,Gau}}$ is assumed to have knowledge of the NN learning parameters and updates after every training step $i \leq k$: learning rate η, weights θ_i, perturbed gradients $\tilde{g}_i$, privacy mechanism $\mathcal{M}_i$, parameters (ϵ, δ), $\mathtt{C}$, the resulting standard deviation σ of the Gaussian distribution and the prior beliefs. The implementation of $\mathcal{A}_{\mathtt{DI}}$ for DPSGD is provided in Algorithm 1.

In each learning step $\mathcal{A}_{\mathtt{DI}}$ first computes the unperturbed, clipped batch gradients for both datasets based on the resulting weights from the previous step of the perturbed learning algorithm (Steps 3 and 4). Then $\mathcal{A}_{\mathtt{DI,Gau}}$ calculates the sensitivity. The ϵ_i and δ_i for each iteration is calculated using RDP composition (cf. Eq. (3)). Consequently, the Gaussian mechanism scale σ is calculated from (ϵ, δ) and Δf using Eq. (1). Using the standard deviation σ, the posterior belief β_i is updated in Step 9 based on the observed perturbed clipped gradient $\tilde{g}_i$ and the unperturbed gradients from Steps 3 and 4. The calculation is based on Lemma 1. After the training finished, $\mathcal{A}_{\mathtt{DI,Gau}}$ tries to identify the used dataset based on the final posterior beliefs β_k on the two datasets. $\mathcal{A}_{\mathtt{DI,Gau}}$ wins the identification game, if $\mathcal{A}_{\mathtt{DI,Gau}}$ chooses the used dataset $\mathcal{D}$. The advantage to win the experiment is statistically estimated from several identical repetitions of the experiment. $\text{Adv}^{\text{DI,Gau}}$ and δ are empirically calculated by counting the cases in which β_k for $\mathcal{D}$ exceeds 0.5 and ρ_β, respectively.

One pass over all records in $\mathcal{D}$ (i.e., one epoch), can comprise multiple update steps. In mini-batch gradient descent, a number of b records from $\mathcal{D}$ is sampled for calculating an update and one epoch results in $|\mathcal{D}|/b$ update steps. In batch gradient descent, all records in $\mathcal{D}$ are used within one update step, and one epoch consists of a single update step. We operate with batch gradient descent, since it reflects the auxiliary side knowledge of $\mathcal{A}_{\mathtt{DI}}$; thus k denotes the overall number of epochs and training steps. In some of the following experiments we will set $\Delta f = LS_{\hat{g}_i}(\mathcal{D})$ in Step 6 by calculating the local sensitivity $LS_{\hat{g}_i}$ for the clipped gradients $\hat{g}_i$ (cf. Definition 3). These assumptions are similar to those of white-box MI attacks against federated learning [33].

Table 1: Time complexity for DS, β and Adv

Algorithm	Complexity	Comment
DS	$O(n^2)$	One-time effort for $\mathcal{D}$.
β	$O(nk)$	Computing belief from clipped Batch gradients.
Adv	$O(1)$	Computing Adv for individual training (cf. 14 in Algorithm 1).

The time complexities for calculating dataset sensitivity, posterior belief and advantage are stated in Table 1. Note that the calculation effort will either lie with $\mathcal{A}_{\mathtt{DI}}$ or the data scientist, depending on whether an audit or an actual attack is performed. The calculation of dataset sensitivity was well parallelizable for the dissimilarity measures considered in this paper.

Algorithm 1 $\mathcal{A}_{\mathtt{DI,Gau}}$ in Deep Learning for Unbounded DP

Require: Neighboring datasets $\mathcal{D}$,$\mathcal{D}'$ with n,n' records, respectively, k, θ_0, η, $\tilde{g}_i$ per training step $i \leq k$, $\mathcal{M}_i$, (ϵ_i, δ_i), prior beliefs $\beta_0(\mathcal{D}) = \beta_0(\mathcal{D}') = 0.5$,

1: **for** $i \in [k]$ **do**
2: **Calculate clipped Batch gradients**
3: $\hat{g}_i(\mathcal{D}) \leftarrow \mathcal{M}_i(\mathcal{D}, \sigma = 0)$
4: $\hat{g}_i(\mathcal{D}') \leftarrow \mathcal{M}_i(\mathcal{D}', \sigma = 0)$
5: **Calculate Sensitivity and** σ
6: $\Delta f \leftarrow GS_{\hat{g}} = \mathtt{C}$
7: $\sigma_i = \Delta f \sqrt{2 \ln(1.25/\delta_i)}/\epsilon_i$
8: **Calculate Belief**
9: $\beta_{i+1}(\mathcal{D}) \leftarrow \frac{\beta_i(\mathcal{D}) \cdot Pr[\mathcal{M}_i(\mathcal{D}, \sigma=\sigma_i) = \tilde{g}_i]}{\beta_i(\mathcal{D}) * Pr[\mathcal{M}_i(\mathcal{D}, \sigma=\sigma_i) = \tilde{g}_i] + \beta_i(\mathcal{D}') \cdot Pr[\mathcal{M}_i(\mathcal{D}') = \tilde{g}_i]}$
10: $\beta_{i+1}(\mathcal{D}') \leftarrow 1 - \beta_{i+1}(\mathcal{D})$
11: **Compute weights**
12: $\theta_{i+1} \leftarrow \theta_i - \eta \tilde{g}_i$
13: **end for**
14: Output $\mathcal{D}$ if $\beta_k(\mathcal{D}) > \beta_k(\mathcal{D}')$, $\mathcal{D}'$ otherwise

5.3 Method for Auditing ϵ

In this section we introduce a method to empirically determine the privacy loss ϵ. This empirical loss is denoted ϵ' and is relevant for data scientists. If ϵ' is close to ϵ, the DP perturbation does not add more noise than necessary. However, if ϵ' is far below ϵ, too much noise is added, and utility is unnecessarily lost. We repeat the training process multiple times and use the set of results to calculate ϵ'. The empirical loss ϵ' can be calculated from different quantities $LS_{\hat{g}}$, β_k, and $\text{Adv}^{\text{DI,Gau}}$ observed during model training:

- From $LS_{\hat{g}_1}, \ldots, LS_{\hat{g}_k}$, the empirical ϵ' is calculated as follows: (i) calculate $\sigma_1, \ldots, \sigma_k$ as $\sigma_i = 2\mathtt{C}/LS_{\hat{g}_i} \cdot \sigma$ (cf. Eq. (2)) for each repetition of the experiment, (ii) calculate ϵ' with RDP composition with target δ, epochs k, and σ using Tensorflow privacy accountant[1], and (iii) choose the maximum value $\epsilon'^{\max}$ over all repetitions of the experiment.

[1]We use Tensorflow Privacy for experiments: https://github.com/tensorflow/privacy.

- From posterior beliefs β, ϵ' is calculated by (i) choosing the maximum final posterior belief $\beta_k^{\max}$ for all experiments and (ii) setting $\epsilon' = \beta_k^{\max}/(1-\beta_k^{\max})$ using Eq. (5).
- From $\text{Adv}^{\text{DI,Gau}}$: (i) counting the number of wins n_{win}, i.e., how often $\beta_k > 0.5$ over all n_{Exp} experiments, (ii) estimate $\text{Adv}^{\text{DI,Gau}} = 2n_{win}/n_{Exp} - 1$, and (iii) calculate $\epsilon' = \sqrt{2\ln(1.25/\delta)}\,\Phi^{-1}\left(\frac{\text{Adv}^{\text{DI,Gau}}+1}{2}\right)$ using Eq. (6).

This empirical loss ϵ' will only be close to ϵ if noise is added according to the sensitivity of the dataset. Of the three variants above, the calculation from the sensitivities is the most direct method. The calculation from the posterior belief is less direct. Since the identification advantage ignores the size of the belief it is expected to be the least accurate way to estimate ϵ.

Furthermore, we also implement the MI adversary $\mathcal{A}_{\texttt{MI}}$ defined by Yeom et al. [47] and compare the resulting advantage to the advantage achieved by $\mathcal{A}_{\texttt{DI,Gau}}$. This instance of $\mathcal{A}_{\texttt{MI}}$ uses the loss of a neural network prediction in an approach similar to $\mathcal{A}_{\texttt{DI,Gau}}$, who analyzes the gradient updates instead.

6 EVALUATION

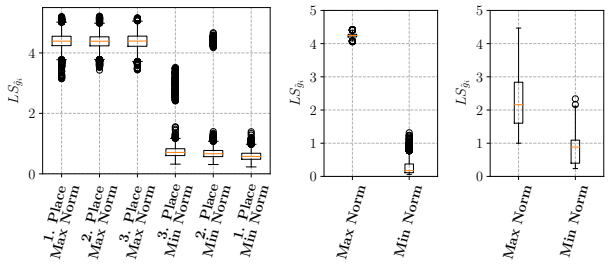
We empirically show that we can train models which yield an empirical privacy loss ϵ' close to the privacy loss bound ϵ. We achieve an advantage equal to ρ_α and tightly bound posterior belief ρ_β when the sensitivity is set to $LS_{\hat{g}_i}$ for the clipped batch gradients at every update step i. Privacy is specified by setting the upper bound for the belief, e.g., to $\rho_\beta = 0.9$. Together with the sensitivity (cf. Section 5.1) this determines the noise of the Gaussian mechanism and yields ϵ. The posterior belief β and the advantage $\text{Adv}^{\text{DI,Gau}}$ are then empirically determined using the implemented adversary $\mathcal{A}_{\texttt{DI,Gau}}$ as described in Section 5.2. The empirical privacy loss ϵ' is determined as described in Section 5.3. We evaluate $\mathcal{A}_{\texttt{DI,Gau}}$ for three ML datasets: the MNIST image dataset[2], the Purchase-100 customer preference dataset [42], and the Adult census income dataset [25]. To improve training speed in our experiments, we set training dataset $\mathcal{D}$ to a randomly sampled subset of size 100 for MNIST and 1000 for both Purchase-100 and Adult. Multiple trainings and perturbations are evaluated on the sampled $\mathcal{D}$.

The MNIST NN consists of two convolutional layers with kernel size $(3,3)$ each, batch normalization and max pooling with pool size $(2,2)$, and a 10-neuron softmax output layer. For Purchase-100, the NN comprises a 600-neuron input layer, a 128-neuron hidden layer and a 100-neuron output layer. Our NN for Adult consists of a 104-neuron input layer due to the use of dummy variables for categorical attributes, two 6-neuron hidden layers and a 2-neuron output layer. We used relu and softmax activation functions for the hidden layers and the output layer. For all experiments we chose the learning rate $\eta = 0.005$ and set the number of iterations $k = 30$ which led to converging models. Preprocessing comprised removal of incomplete records, and data normalization.

6.1 Evaluation of Sensitivities

While local sensitivity is favored when striving to reach the privacy bound, we evaluate and compute both $\Delta f = \hat{LS}_{\hat{g}_i}(\mathcal{D})$ and $\Delta f = GS_{\hat{g}}$, as described in Section 5.1. In addition, we consider

[2]Dataset and detailed description available at: http://yann.lecun.com/exdb/mnist/



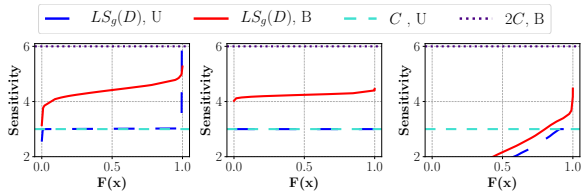
(a) MNIST (b) Purchase-100 (c) Adult

Figure 4: Distribution of the local sensitivity $LS_{\hat{g}_i}(\mathcal{D})$ computed by the adversary using Eq. (9) from max to min difference in $\mathcal{D}$ and $\mathcal{D}'$ for all 30 epochs i, repeated 250 times

bounded and unbounded DP in our experiments. In order to find the most dissimilar data point for the construction of $\hat{\mathcal{D}}'$ in Eq. (9) and Eq. (11) we require a dissimilarity measure. We considered domain specific candidates for the dissimilarity measures: the negative structural similarity index measure (SSIM) and Euclidean distance for MNIST, and the Hamming, Euclidean, Manhattan, and Cosine distance for the datasets Purchase-100 and Adult. We chose these metrics because we expect them to contain information relevant to the gradients of data points. However, for example we quickly noticed for the Euclidean distance on MNIST image data that it does not capture the meaning or shapes pictured and thus falls short. Instead, the SSIM captures structure in images, and images with a small SSIM dissimilarity values resulted in similar gradients, while images with greater dissimilarity resulted in very different gradients. This observation supports the hypothesis that an appropriate domain-specific measure can be used to estimate local sensitivity $LS_{\hat{g}_i}$ from dataset sensitivity DS. For Purchases-100 the Hamming distance was superior, and for Adult the Manhattan distance worked best. For the sensitivity experiments the bound for the posterior belief is set to $\rho_\beta = 0.9$. Each experiment concerning dataset sensitivity is repeated $n_{Exp} = 250$ times.

To confirm that maximizing dataset sensitivity from Definition 6 allows us to approximate $LS_{\hat{g}_i}$, we train with several differing $\mathcal{D}'$ and evaluate the sensitivities for all $k = 30$ iterations. For the MNIST dataset, the top three choices of $\mathcal{D}'$ that maximize DS and the three choices that minimize DS are used. As expected, the resulting local sensitivities $LS_{\hat{g}_i}$ shown in Figure 4a are clearly larger for the three top choices. The outliers for the second and third smallest dataset sensitivities only account for 1.6% and 5.2% of the 7500 overall observed sensitivity norms. More importantly, no far outliers occur for the largest and smallest sensitivities. The same general trend holds for Purchase-100 and Adult in Figures 4b and 4c, which we limit to the maximum and minimum DS due to space constraints.

If the chosen global sensitivity is too large compared to the local sensitivity of a specific dataset too much noise will be added when using $GS_{\hat{g}}$, as described in Section 5.1. Global sensitivity $GS_{\hat{g}}$ and local sensitivity $LS_{\hat{g}_i}$ are determined for bounded and unbounded DP over $n_{Exp} = 1000$ repetitions for $\rho_\beta = 0.9$ ($\epsilon = 2.2$) according to Eq. (10) and Eq. (12). They can be compared in Figure 5.



(a) MNIST (b) Purchase-100 (c) Adult

Figure 5: Sensitivities over the course of the training for $\rho_\beta = 0.9$ ($\epsilon = 2.2$) and $\mathtt{C} = 3$

Table 2: Empirical $\text{Adv}^{\text{DI,Gau}}$ and δ' for $\rho_\beta = 0.9$ using $LS_{\hat{g}_i}$ and $GS_{\hat{g}}$ with bounded (B) and unbounded (U) DP

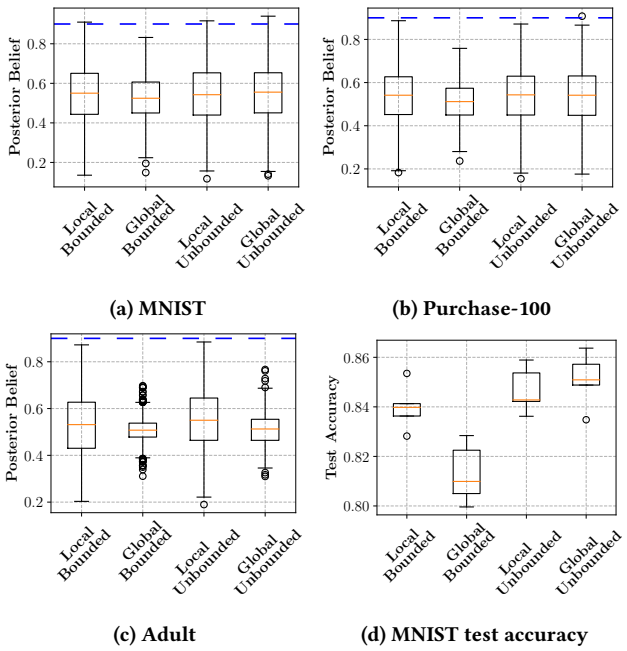
	MNIST		Purchase-100		Adult	
	$\text{Adv}^{\text{DI,Gau}}$	δ'	$\text{Adv}^{\text{DI,Gau}}$	δ'	$\text{Adv}^{\text{DI,Gau}}$	δ'
LS B	0.24	2e-3	0.25	0	0.17	0
LS U	0.23	2e-3	0.23	0	0.22	0
GS B	0.18	0	0.1	0	0.13	0
GS U	0.27	4e-3	0.24	1e-3	0.18	0

6.2 Quantification of Identifiability for DPSGD

For each of the 1000 experiment repetitions, the posterior belief β_k and the membership advantage $\text{Adv}^{\text{DI,Gau}}$ are experimentally determined using the implementation of $\mathcal{A}_{\text{DI,Gau}}$ for DPSGD. We set $\rho_\beta = 0.9$ ($\epsilon = 2.2$) and compare bounded and unbounded DP. Table 2 shows the analytically obtained values for privacy loss ϵ, and the bound ρ_α for the advantage. The parameters ϵ, δ, and ρ_α for $\rho_\beta = 0.9$ can be read from Table 3; ϵ is determined from Eq. (5), whereas ρ_α is calculated from ϵ from Theorem 3.

First, we verify that the upper bound ρ_β on the posterior belief holds. The posterior beliefs β_k of these experiments are described in Figures 6a, 6b and 6c. For a single experiment the posterior belief on the training dataset $\mathcal{D}$ is on average only slightly above 0.5. While for most cases the posterior belief is far below the bound of 0.9 (specified by the blue, dashed line), the upper bound is violated with a small probability. The relative frequency of these violations is denoted as δ'. Since the DP bound, and thus ρ_β, only holds with probability $1 - \delta$ according to Theorem 2 violations are acceptable as long as $\delta' \leq \delta$. Indeed, the experimentally obtained δ' for $\rho_\beta = 0.9$ in Table 2 is always smaller than the corresponding δ in Table 3. Similarly, the advantage should be close to the estimate ρ_α stated in Table 3. The advantage is experimentally estimated as the relative frequency of experiments where the implemented adversary $\mathcal{A}_{\text{DI,Gau}}$ correctly chooses $\mathcal{D}$ and is stated in Table 2.

Figure 6 illustrates the influence of sensitivity in the bounded and unbounded DP settings. In Figures 6a, 6b and 6c, the chosen upper bound $\rho_\beta = 0.9$ (blue line) is clearly not reached for the bounded case when global sensitivities are used. Similarly, the advantage of $\mathcal{A}_{\text{DI,Gau}}$ in Table 2 is smaller when the global sensitivity is used. Here it holds that $LS_{\hat{g}_i}(\mathcal{D}) < 2\mathtt{C} = \Delta f$, which implies that the examples differing between $\mathcal{D}'$ and $\mathcal{D}$ do not point in opposite directions in the bounded setting. For the unbounded DP case, this effect is not observed with the MNIST and Purchase-100 datasets.



(a) MNIST (b) Purchase-100

(c) Adult (d) MNIST test accuracy

Figure 6: Distribution of empirical posterior beliefs β_k (panels a to c) and an example for test accuracy after training with $\rho_\beta = 0.9$ ($\epsilon = 2.2$), for local and global sensitivity, bounded and unbounded DP (panel d)

Instead, the use of local and global sensitivity leads to the same distribution of posterior beliefs and approximately the same advantage. This result stems from the fact that the per-example gradients over the course of all epochs were close to or greater than $\mathtt{C} = 3$, i.e., the differentiating example in $\mathcal{D}$ must have the gradient magnitude $\mathtt{C} = 3$. However, in the Adult dataset, $LS_{\hat{g}_i}(\mathcal{D}) < \mathtt{C} = 3$, so too much noise is added using $GS_{\hat{g}}$ in the unbounded DP setting as well.

From a practical standpoint, these observations are critical, since unnecessary noise degrades the utility of the model when the global sensitivity is too large, as shown in Figure 6d. While all experiments were done with $\mathtt{C} = 3$, we expect a similar relationship between $LS_{\hat{g}_i}$ and $GS_{\hat{g}}$ for different values of $\mathtt{C}$, since we observed the unclipped gradients to usually be greater than $\mathtt{C} = 3$.

6.3 Auditing DPSGD

This section details the audit of ϵ. As shown in Section 5.3, the calculation of the empirical loss ϵ' can be based on (i) the local sensitivity, (ii) the posterior beliefs β_k or (iii) on the advantage $\text{Adv}^{\text{DI,Gau}}$. To validate that ϵ' is close to the target privacy loss ϵ we use the setting described in Section 6.2 and Table 3.

The resulting empirical loss ϵ' is compared to the the target privacy loss ϵ for the bounded case in Figures 7 to 9. As expected Figures 7a, 8a and 9a support that the privacy loss ϵ can be best estimated from the local sensitivity: the red curve lies on the ideal green curve. The estimation is less precise from the posterior beliefs and shown in Figures 7b, 8b and 9b. The estimation is worst from the advantage in Figures 7c, 8c and 9c, where the red curve deviates most from the ideal green curve for all datasets. It is evident that

Table 3: Experiment setting for posterior belief ρ_β and δ with analytically determined privacy loss ϵ and advantage bound ρ_α

	MNIST				Purchase-100				Adult			
ρ_β	0.52	0.75	0.9	0.99	0.53	0.75	0.9	0.99	0.53	0.75	0.9	0.99
δ	0.01				0.001				0.001			
ϵ	0.08	1.1	2.2	4.6	0.12	1.1	2.2	4.6	0.12	1.1	2.2	4.6
ρ_α	0.01	0.14	0.28	0.54	0.01	0.12	0.23	0.46	0.01	0.12	0.23	0.46

the use of global sensitivity (blue lines) results in an underestimation of ϵ for all datasets. When local sensitivity is used, the small deviation from the ideal curve confirms that $\mathcal{A}_{\mathtt{DI,Gau}}$ comes close to the theoretical privacy guarantees offered by DP. A data scientist who specifies ϵ via the identifiability bounds ρ_α and ρ_β can audit ϵ using the implementation of $\mathcal{A}_{\mathtt{DI,Gau}}$. We see that in some cases $\epsilon' > \epsilon$, or equivalently $\beta_k(\mathcal{D}) > \rho_\beta$. These variations are due to the probabilistic nature of the estimation and the bound only holds with probability 1-δ. Furthermore, we observe in some occasions that $\text{Adv}^{\text{DI,Gau}} > \rho_\alpha$ which stems from the fact that $\text{Adv}^{\text{DI,Gau}}$ is an expected value for a series of experiments, which falls within a confidence interval around ρ_α.

To enable comparison with membership inference we implemented $\mathcal{A}_{\mathtt{MI}}$ by expanding the implementation of Jayaraman and Evans [20], which implements the attack suggested by Yeom et al. [47]. Figures 7d, 8d, and 9d visualize the advantage resulting from both $\mathcal{A}_{\mathtt{DI,Gau}}$ and $\mathcal{A}_{\mathtt{MI}}$ for our setting, as well as the bounds provided by the DP guarantee and the MI bound of Yeom et al. [47]. We see that the MI bound is very loose for all evaluated datasets, as previously noted by Jayaraman and Evans [20]. Furthermore, we see that our implementation of $\mathcal{A}_{\mathtt{DI,Gau}}$ significantly outperforms $\mathcal{A}_{\mathtt{MI}}$ on all datasets and values of ϵ.

7 DISCUSSION

$\mathcal{A}_{\mathtt{DI}}$ diverges from other attacks against DP or ML, which necessitates a discussion of $\mathcal{A}_{\mathtt{DI}}$'s properties in relation to alternative approaches. Our goal is to construct an adversary that most closely challenges DP, and can be connected to societal norms and legislation via identifiability score. To this end, $\mathcal{A}_{\mathtt{DI}}$ has knowledge of all but one element of the training data and the gradients at every update step. Since the DP guarantee must hold in the presence of all auxiliary information, both of these assumptions relate the attack model $\mathcal{A}_{\mathtt{DI}}$ directly to the DP guarantee. Since $\mathcal{A}_{\mathtt{DI}}$ has knowledge of all but one element instead of only the distribution, $\mathcal{A}_{\mathtt{DI}}$ possesses significantly more information than MI adversaries.

A natural question arises w.r.t. $\mathcal{A}_{\mathtt{DI}}$'s practical relevance. Especially in a federated learning setting $\mathcal{A}_{\mathtt{DI}}$ knows the gradients during every update step, if participating as a data owner. Furthermore, $\mathcal{A}_{\mathtt{DI}}$ could realistically obtain knowledge of a significant portion of the training data, since public reference data is often used in training datasets and only extended with some custom training data records, necessitating the notion of DP in general.

To further comment on the utility that can be achieved from a differentially private model, we note that the optimal choice for $\mathtt{C}$ may stray from the original recommendation of Abadi et al. [1]. We follow this recommendation and set $\mathtt{C} = 3$, which limits the utility loss that results when $\mathtt{C}$ is too large (unnecessary noise addition) and too small (loss of information about the gradient). Since this balance holds for unbounded DP and does not consider the notion of local sensitivity, we expect that a different $\mathtt{C}$ may yield better utility than what we report. Varying $\mathtt{C}$ may also change the balance between local sensitivity and global sensitivity from Figures 7 to 9. Furthermore, since gradients change over the course of training, the optimal value of $\mathtt{C}$ at the beginning of training may no longer be optimal toward the end of training according to McMahan et al. [30]. Adaptively setting the clipping norm as suggested by Thakkar et al. [44] may improve utility by changing $\mathtt{C}$ as training progresses. We expect that doing so might bring ϵ' closer to ϵ when auditing the DP guarantee, and achieve similar by using local sensitivity.

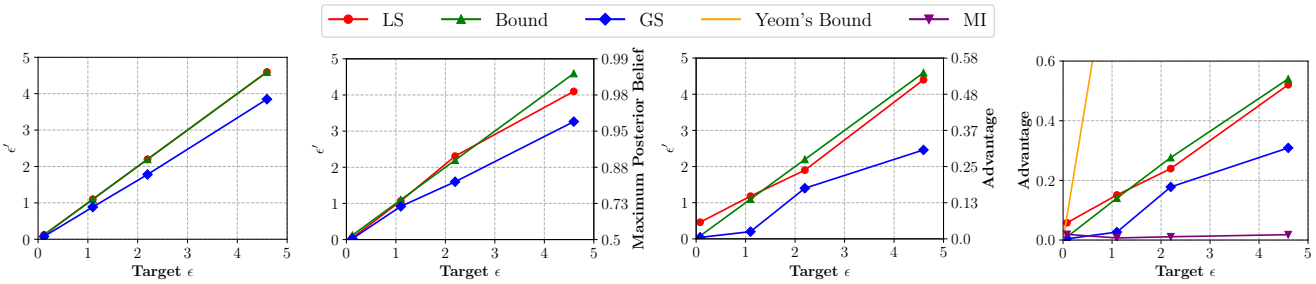
8 RELATED WORK

Choosing and interpreting DP privacy parameters has been addressed from several directions.

Lee and Clifton [26, 27] proposed DI as a Bayesian privacy notion which quantifies ϵ w.r.t. an adversary's maximum posterior belief ρ_β on a finite set of possible input datasets. Yet, both papers focus on the scalar ϵ Laplace mechanism without composition, while we consider the (ϵ, δ) multidimensional Gauss mechanism under RDP composition. Li et al. [28] demonstrate that DI matches the DP definition when an adversary decides between two neighboring datasets $\mathcal{D}, \mathcal{D}'$. Kasiviswanathan et al. [24] also provide a Bayesian interpretation of DP. While they also formulate posterior belief bounds and discuss local sensitivity, they do not cover expected advantage and implementation aspects such as dataset sensitivity.

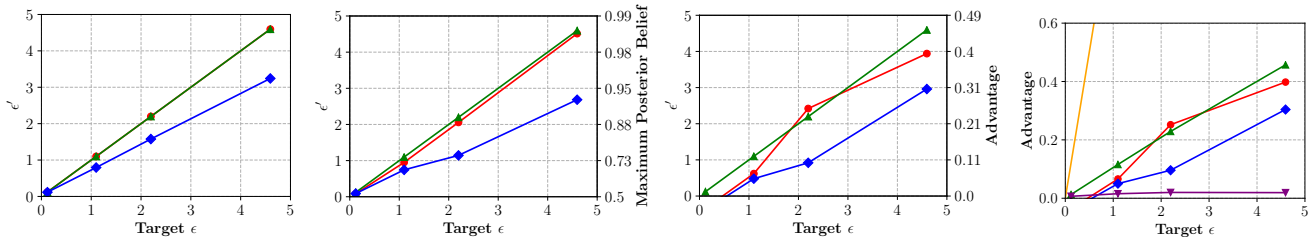
The choice of privacy parameter ϵ has been tied to economic consequences. Hsu et al. [17] derive a value for ϵ from a probability distribution over a set of negative events and the cost for compensation of affected participants. Our approach avoids the ambiguity of selecting bad events. Abowd and Schmutte [2] describe a social choice framework for choosing ϵ, which uses the production possibility frontier of the model and the social willingness to accept privacy and accuracy loss. We part from their work by choosing ϵ w.r.t. the advantage of the DP adversary. Eibl et al. [14] propose a scheme that allows energy providers and consumers to negotiate DP parameters by fixing a tolerable noise scale of the Laplace mechanism. The noise scale is transformed into the individual posterior belief of the DP adversary per energy consumer. We part from their individual posterior belief analysis and use the local sensitivity between two datasets chosen by the dataset sensitivity heuristic.

The evaluation of DP in a deep learning setting has largely focused on MI attacks [5, 6, 16, 20, 21, 40, 42]. From Yeom et al. [47] we take the idea of bounding membership advantage in terms of DP privacy parameter ϵ. However, while MI attacks evaluate the DP privacy parameters in practice, DP is defined to offer protection from far stronger adversaries, as Jayaraman et al. [20] empirically validated. Humphries et al. [18] derive a bound for membership



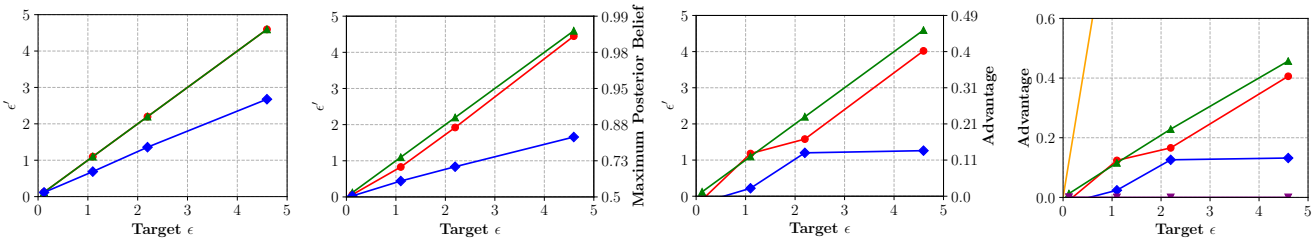
(a) ϵ' from $\Delta f_0, \ldots, \Delta f_k$ (b) ϵ' from posterior belief β_k (c) ϵ' from advantage $\text{Adv}^{\text{DI,Gau}}$ (d) Comparison to MI

Figure 7: Audit of ϵ (a-c) and comparison with $\mathcal{A}_{\text{MI}}$ (d) for MNIST data (bounded case)



(a) ϵ' from $\Delta f_0, \ldots, \Delta f_k$ (b) ϵ' from posterior belief β_k (c) ϵ' from advantage $\text{Adv}^{\text{DI,Gau}}$ (d) Comparison to MI

Figure 8: Audit of ϵ (a-c) and comparison with $\mathcal{A}_{\text{MI}}$ (d) for Purchase-100 data (bounded case)



(a) ϵ' from $\Delta f_0, \ldots, \Delta f_k$ (b) ϵ' from posterior belief β_k (c) ϵ' from advantage $\text{Adv}^{\text{DI,Gau}}$ (d) Comparison to MI

Figure 9: Audit of ϵ (a-c) and comparison with $\mathcal{A}_{\text{MI}}$ (d) for Adult data (bounded case)

advantage that is tighter than the bound derived by Yeom et al. [47] by analyzing an adversary with additional information. Their work analyzes the impact of giving up the i.i.d. assumption and does not suggest an implementation of the DP adversary. Jagielski et al. [19] estimate empirical privacy guarantees based on Monte Carlo approximations. While they use active poisoning attacks to construct datasets $\mathcal{D}$ and $\mathcal{D}'$ that result in maximally different gradients under gradient clipping, we define dataset sensitivity, which does not require the introduction of malicious samples.

9 CONCLUSION

We defined two identifiability bounds for the DP adversary in ML with DPSGD: maximum posterior belief ρ_β and expected membership advantage ρ_α. These bounds can be transformed to privacy parameter ϵ. In consequence, with ρ_α and ρ_β, data owners and data scientists can map legal and societal expectations w.r.t. identifiability to corresponding DP privacy parameters. Furthermore, we implemented an instance of the DP adversary for ML with DPSGD and showed that it allows us to audit parameter ϵ. We evaluated the effect of sensitivity in DPSGD and showed that our upper bounds are reached under composition. To reach the bounds and thus improve utility the sensitivity must reflect the local sensitivity of the training dataset which we approximate for DPSGD with a heuristic.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for suggesting numerous improvements. This work has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 825333 (MOSAICrOWN), and the Federal State of Salzburg under the WISS2025 program. We used datasets from the UCI machine learning repository [8].

REFERENCES

[1] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the Conference on Computer and Communications Security (CCS)*. ACM Press, New York, NY, USA, 308–318.

[2] John M. Abowd and Ian M. Schmutte. 2019. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review* 109, 1 (January 2019), 171–202.

[3] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private Empirical Risk Minimization. In *Proceedings of Symposium on Foundations of Computer Science (SFCS)*. IEEE Computer Society, Piscataway, NJ, USA, 464–473.

[4] Daniel Bernau, Günther Eibl, Philip W. Grassal, Hannah Keller, and Florian Kerschbaum. 2021. Quantifying identifiability to choose and audit ϵ in differentially private deep learning. arXiv:2103.02913 [cs.CR]

[5] Daniel Bernau, Jonas Robl, Philip W. Grassal, Steffen Schneider, and Florian Kerschbaum. 2021. Comparing Local and Central Differential Privacy Using Membership Inference Attacks. In *Proceedings of the Conference on Data and Applications Security and Privacy (DBSEC)*. Springer International Publishing, Cham, DEU, 22–42.

[6] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In *Proceedings of the Conference on Computer and Communications Security (CCS)*. ACM Press, New York, NY, USA, 343–362.

[7] Chris Clifton and Tamir Tassa. 2013. On syntactic anonymity and differential privacy. In *Proceedings of the Conference on Data Engineering Workshops (ICDEW)*. IEEE Computer Society, Piscataway, NJ, USA, 88–93.

[8] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. Retrieved September 2, 2021 from https://archive.ics.uci.edu/ml

[9] Cynthia Dwork. 2006. Differential Privacy. In *Proceedings of the Colloquium on Automata, Languages and Programming (ICALP)*. Springer-Verlag, Berlin, Heidelberg, DEU, 1–12.

[10] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. 2006. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Proceedings of the Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer-Verlag, Berlin, Heidelberg, DEU, 486–503.

[11] Cynthia Dwork and Aaron Roth. 2014. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (August 2014), 211–407.

[12] Cynthia Dwork and Guy N. Rothblum. 2016. Concentrated Differential Privacy. arXiv:1603.01887 [cs.DS]

[13] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. 2010. Boosting and Differential Privacy. In *Proceedings of Symposium on Foundations of Computer Science (SFCS)*. IEEE Computer Society, Piscataway, NJ, USA, 51–60.

[14] Günther Eibl, Kaibin Bao, Philip-William Grassal, Daniel Bernau, and Hartmut Schmeck. 2018. The influence of differential privacy on short term electric load forecasting. *Energy Informatics* 1, 1 (October 2018), 93–113.

[15] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. In *Proceedings of the Conference on Learning Representations (ICLR)*. IEEE Computer Society, Piscataway, NJ, USA.

[16] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership Inference Attacks Against Generative Models. In *Proceedings of the Privacy Enhancing Technologies Symposium (PETS)*. Sciendo, Warsaw, POL, 133–152.

[17] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin Pierce, and Aaron Roth. 2014. Differential Privacy: An Economic Method for Choosing Epsilon. In *Proceedings of the Computer Security Foundations Workshop (CSFW)*. IEEE Computer Society, Piscataway, NJ, USA, 398–410.

[18] Thomas Humphries, Matthew Rafuse, Lindsey Tulloch, Simon Oya, Ian Goldberg, and Florian Kerschbaum. 2020. Differentially Private Learning Does Not Bound Membership Inference. arXiv:2010.12112 [cs.CR]

[19] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing Differentially Private Machine Learning: How Private is Private SGD?. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates Inc., Red Hook, NY, USA.

[20] Bargav Jayaraman and David Evans. 2019. Evaluating Differentially Private Machine Learning in Practice. In *Proceedings of the USENIX Security Symposium*. USENIX Association, Berkeley, CA, USA, 1895–1912.

[21] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. 2020. Revisiting Membership Inference Under Realistic Assumptions. arXiv:2005.10881 [cs.CR]

[22] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2015. The Composition Theorem for Differential Privacy. In *Proceedings of the Conference on Machine Learning (ICML)*. PMLR, 1376–1385.

[23] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. 2017. The Composition Theorem for Differential Privacy. *IEEE Transactions on Information Theory* 63, 6 (2017), 4037–4049.

[24] Shiva P. Kasiviswanathan and Adam Smith. 2014. On the Semantics of Differential Privacy: A Bayesian Formulation. *Journal on Privacy and Confidentiality* 6 (2014), 1–16. Issue 1.

[25] Ron Kohavi. 1996. Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*. AAAI Press, Palo Alto, CA, USA, 202–207.

[26] Jaewoo Lee and Chris Clifton. 2011. How Much is Enough? Choosing Epsilon for Differential Privacy. In *Proceedings of the Conference on Information Security (ISC)*. Springer-Verlag, Berlin, Heidelberg, DEU, 325–340.

[27] Jaewoo Lee and Chris Clifton. 2012. Differential Identifiability. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*. ACM Press, New York, NY, USA, 1041–1049.

[28] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. 2013. Membership Privacy: A Unifying Framework for Privacy Definitions. In *Proceedings of the Conference on Computer and Communications Security (CCS)*. ACM Press, New York, NY, USA, 889–900.

[29] Kantilal Vardichand Mardia, John T. Kent, and John M. Bibby. 1979. *Multivariate analysis*. Academic Press, New York, NY, USA.

[30] Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. 2018. Learning Differentially Private Recurrent Language Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.

[31] H. Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. 2019. A General Approach to Adding Differential Privacy to Iterative Training Procedures. arXiv:1812.06210 [cs.LG]

[32] Ilya Mivonov. 2017. Rényi Differential Privacy. In *Proceedings of the Computer Security Foundations Symposium (CSF)*. IEEE Computer Society, Piscataway, NJ, USA, 263–275.

[33] Milad Nasr, Reza Shokri, and Amir Houmansadr. 2019. Comprehensive Privacy Analysis of Deep Learning: Stand-alone and Federated Learning under Passive and Active White-box Inference Attacks. In *Proceedings of the Symposium on Security and Privacy (S&P)*. IEEE Computer Society, Piscataway, USA, 739–753.

[34] Helen Nissenbaum. 2016. Differential Privacy in Context: Conceptual and Ethical Considerations. In *Four Facets of Differential Privacy Symposium*. Presented at the Institute for Advanced Study, Princeton, NJ, USA. Retrieved September 2, 2021 from https://www.ias.edu/ideas/2016/differential-privacy-symposium

[35] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. 2007. Smooth Sensitivity and Sampling in Private Data Analysis. In *Proceedings of the Symposium on Theory of Computing (STOC)*. ACM Press, New York, NY, USA, 75–84.

[36] Kobbi Nissim and Alexandra Wood. 2018. Is privacy privacy? *Philosophical Transactions of the Royal Society* 376, 2128 (2018).

[37] American Department of Health and Human Services. 2010. Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Retrieved September 2, 2021 from https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

[38] European Parliament and Council of the European Union. 2016. General Data Protection Regulation. *Official Journal of the European Union L119* 59, 1 (May 2016), 1–88.

[39] Article 29 Data Protection Working Party. 2014. Opinion 05/2014 on Anonymisation Techniques. Retrieved September 2, 2021 from https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

[40] Md. Atiqur Rahman, Tanzila Rahman, Robert Laganière, and Noman Mohammed. 2018. Membership Inference Attack against Differentially Private Deep Learning Model. *Transactions on Data Privacy* 11 (2018), 61–79.

[41] Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving Deep Learning. In *Proceedings of the Conference on Computer and Communication Security (CCS)*. ACM Press, New York, NY, USA, 1310–1321.

[42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership Inference Attacks against Machine Learning Models. In *Proceedings of the Symposium on Security and Privacy (S&P)*. IEEE Computer Society, Piscataway, NJ, USA, 3–18.

[43] Shuang Song, Kamalika Chaudhuri, and Anand D. Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *Proceedings of the Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE Computer Society, Piscataway, NJ, USA, 245–248.

[44] Om Thakkar, Galen Andrew, and H. Brendan McMahan. 2019. Differentially Private Learning with Adaptive Clipping. arXiv:1905.03871 [cs.LG]

[45] Kagan Tumer and Joydeep Ghosh. 1996. Estimating the Bayes error rate through classifier combining. In *Proceedings of the Conference on Pattern Recognition (ICPR)*. IEEE Computer Society, Piscataway, NJ, USA, 695–699.

[46] Tim van Erven and Peter Harremoës. 2010. Rényi Divergence and Majorization. In *Proceedings of the Symposium on Information Theory (ISIT)*. IEEE Computer Society, Piscataway, NJ, USA, 1335–1339.

[47] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In *Proceedings of the Computer Security Foundations Symposium (CSF)*. IEEE Computer Society, Piscataway, NJ, USA, 268–282.