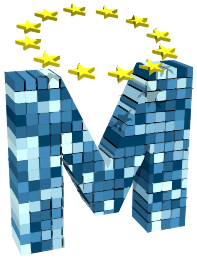


Project title: Multi-Owner data Sharing for Analytics and Integration respecting Confidentiality and OWNer control
Project acronym: MOSAICrOWN
Funding scheme: H2020-ICT-2018-2
Topic: ICT-13-2018-2019
Project duration: January 2019 – December 2021

D5.3
Final Report on
Privacy Metrics, Risks, and Utility

Editors: Daniel Bernau (SAP SE)
Reviewers: Stefano Paraboschi (UNIBG)
Pierangela Samarati (UNIMI)

Abstract

Research work in WP5 advances the state of the art in sanitization techniques, considering approaches for the anonymization of data based on k-anonymity and differential privacy. One goal of WP5 is to provide techniques for scalable use of sanitization techniques in data markets (i.e., by data owners and data scientists, in addition to privacy experts). This deliverable focuses on the use of metrics for the improved application of differential privacy in a Machine Learning scenario. To this end, we see high potential in privacy metrics that wrap anonymization parameters and relate to lawmaker requirements, and ease parameter setting for data scientists and data owners. In this deliverable we will illustrate how differential privacy anonymization parameters (ε, δ) can be transformed into identifiability metrics, especially in machine learning. The use of differential privacy during the training phase of a machine learning model offers a scalable anonymization technique. However, selecting the privacy parameters for differential privacy is a challenging task for data scientists since the actual strength of the parameters is dataset dependent. One potential remedy is to derive privacy parameters from semantic metrics. Here, membership inference attacks have received a lot of attention in the context of machine learning. However, membership inference attacks are strictly weaker than the attacks against which differential privacy protects, and thus privacy parameters chosen under membership inference will likely be too high. We formulate two identifiability bounds for the differential privacy adversary and show that these bounds can actually be reached. We are optimistic that these bounds can support data scientists in choosing privacy parameters, and that the bounds derive more efficient privacy parameters in comparison to previous work. In comparison to using membership inference attacks for measuring the strength of privacy parameters, our bounds are for the strong adversary assumed by DP and thus almost tight.

Type	Identifier	Dissemination	Date
Deliverable	D5.3	Public	2020.12.31

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825333.

MOSAICrOWN Consortium

1.	Università degli Studi di Milano	UNIMI	Italy
2.	EMC Information Systems International	EISI	Ireland
3.	Mastercard Europe	MC	Belgium
4.	SAP SE	SAP SE	Germany
5.	Università degli Studi di Bergamo	UNIBG	Italy
6.	GEIE ERCIM (Host of the W3C)	W3C	France

Versions

Version	Date	Description
0.1	2020.11.27	Initial Release
0.2	2020.12.18	Second Release
1.0	2020.12.31	Final Release

List of Contributors

This document contains contributions from different MOSAICrOWN partners. Contributors for the chapters of this deliverable are presented in the following table.

Chapter	Author(s)
Executive Summary	Daniel Bernau (SAP SE)
Chapter 1: Introduction	Daniel Bernau (SAP SE)
Chapter 2: Preliminaries	Daniel Bernau (SAP SE)
Chapter 3: Identifiability bounds for the differential privacy adversary in Machine Learning	Daniel Bernau (SAP SE)
Chapter 4: Derivation of tight upper bounds	Daniel Bernau (SAP SE)
Chapter 5: Application to deep learning using DPSGD	Daniel Bernau (SAP SE)
Chapter 6: Conclusions	Daniel Bernau (SAP SE)

Contents

List of Figures

Executive Summary

Differential privacy allows data scientists to bound the influence that members in the training data have on a machine learning model. To use differential privacy in machine learning with the differentially private stochastic gradient descent, data scientists must choose privacy parameters, e.g. (ε, δ) for the Gaussian mechanism. Values for (ε, δ) are difficult to choose because these worst-case upper bounds might not be tight for practical datasets. Concrete membership inference attacks have been used to choose ε, but represent an empirically observed lower bound. Differential privacy aims to protect against adversaries with arbitrary auxiliary information, so only adversaries stronger than the membership inference adversary can lead to empirically verifiable upper bounds. Furthermore, the privacy parameters (ε, δ) do not match societal norms and legal requirements w.r.t. factual identifiability of the underlying training data when receiving a classification from a differentially private machine learning model.

We put forward the idea of inferring privacy parameters (ε, δ) based on an adversary's Bayesian belief about the presence of a specific record in the training dataset. We bound the posterior belief for multidimensional queries under advanced composition, and observe that this bound can actually be tight in practice. Furthermore, we connect the strong adversary considered by differential privacy to membership inference bounds by deriving a membership advantage bound. Posterior belief and expected membership advantage are derived directly from the differential privacy definition and protect against the strong adversary with arbitrary auxiliary knowledge about all but one record in a database. Therefore, data owners can choose (ε, δ) based on two tight identifiability-based metrics, namely *maximum posterior belief* and *expected membership advantage*.

1. Introduction

The research work in WP5 advances the state of the art in sanitization techniques, considering approaches for the anonymization of data based on k-anonymity and differential privacy. One goal of WP5 is to provide techniques for scalable use of sanitization techniques in data markets (i.e., by data owners and data scientists, in addition to privacy experts). This deliverable focuses on the use of metrics for the improved application of differential privacy in a Machine Learning scenario.

Differential privacy (DP) has received much attention by the privacy research community, leading to key contributions such as the tight estimation of privacy loss under composition [Miv17] and mechanisms for differentially private deep learning [ACG+16]. However, data scientists have to choose privacy parameter (ε, δ) for training a machine learning model with the differentially private stochastic gradient descent. Several approaches for choosing and interpreting privacy parameters have been introduced, yet they do not reflect identifiability [AS19, HGH+14], part from the original differential privacy definition [YGFJ18, LC12, RRLM18, BGRK19] or lack applicability to common differential privacy mechanisms [LC11]. Especially in deep learning, practical membership inference (MI) attacks have been used to measure identifiability (e.g.,[BGRK19, RRLM18, HMDD19, JE19, JWK+20, CYZF20, SSSS17, YGFJ18]). However, membership inference adversaries do not have unconstrained auxiliary information. DP adversaries are assumed to possess unconstrained auxiliary information and DP thus provides an upper bound on a strong adversary with background information about all data points in the input data but one. Therefore, MI attacks offer intuition about the outcome of practical attackers, but bounds on MI attacks in terms of differential privacy ε are not tight [JE19], and thus MI can only represent a lower bound on identifiability.

Furthermore, since differential privacy aims to encourage data owners to participate in research studies that benefit society, intuitive communication of risk to individuals will strongly affect its widespread implementation [Nis16]. Data owners will likely only agree to offer their data to a research study or as training data for a machine learning model, if they can confidently assess their privacy protection. However, the factual identifiability risk is only indirectly specified by DP privacy parameters (ε, δ) [NW18]. In fact, if data owners are told their answers to a survey are guaranteed to change the outcome only very slightly (i.e., by a factor e^{ε}) participation in the survey may even decrease, since data owners feel their contribution is not important [OK20]. Furthermore, some privacy regulations refer to individual identifiability [Eur16, Ame10] as a measure for anonymization, a concept that cannot directly be mapped to DP privacy parameters (e.g., Clifton et al. [CT13]). In consequence, if DP is used to comply with privacy regulations [NW18, Par14] interpreting the factual guarantees w.r.t. identifiability risk for privacy parameters (ε, δ) is required.

We formulate identifiability bounds and transform these bounds into concrete privacy parameters (ε, δ). Rather than analyzing the MI adversary, we consider a Differential Identifiability adversary with unconstrained auxiliary knowledge and derive the maximum *posterior belief* ρ_β of the adversary as a worst case Bayesian identifiability bound related to (ε, δ). Furthermore,

we define the complementary metric of *expected membership advantage* ρ_a. Expected membership advantage offers a quantification of the adversary's probability of correctly identifying the underlying dataset $\mathcal{D}$ or $\mathcal{D}'$. Concretely, the expected membership advantage specifies how often posterior belief is greater than 50% and allows us a direct comparison with the membership advantage of Yeom et al. [YGFJ18] for the membership inference adversary.

A subsequent question is whether our identifiability bounds are actually tight. Holding the distribution of differential privacy noise addition constant, multiple global (ε, δ) guarantees will result from various chosen sensitivities, which quantify the difference between data sets that is covered by noise [NRS07]. Therefore, the factual guarantee (ε, δ) depends on the difference between data sets and the identifiability bounds might accordingly not be tight. In differentially private deep learning, noise is scaled to the difference between possible gradients; however, the estimated global sensitivity might far exceed the factual sensitivity, since the training data records are likely to be within the same domain (e.g., pictures of cars vs. pictures of nature scenes). We propose scaling the sensitivity to the difference between the gradients of a fixed data set and any neighboring dataset to achieve a tight bound. We evaluate how tight our identifiability bounds are for one data analytics and one machine learning reference data set, and show that we can indeed achieve tight bounds. Our main contributions are:

- Formulate identifiability bounds for the posterior belief and the expected membership advantage that can be transformed into privacy parameters (ε, δ) and used in conjunction with composition.
- The practical implementation of an adversary that meets all assumptions of worst-case adversaries against DP.
- A heuristic for scaling sensitivity of the differentially private stochastic gradient descent. This heuristic leads to tight bounds.

This deliverable is structured as follows. Preliminaries are presented in Chapter 2, where we provide an overview of notations and concepts that are used throughout this deliverable. Afterwards, we formulate identifiability metrics and analyze their upper bounds in Chapter 3 and Chapter 4. The metrics are evaluated for a reference dataset from the deep learning domain in Chapter 5. We present conclusions in Chapter 6.

2. Preliminaries

In the following three sections we will present the building blocks that are used and extended in this deliverable. Section 2.1 first provides fundamentals with respect to differential privacy, mechanisms for differentially private machine learning and composition. Secondly, Section 2.2 and Section 2.3 introduce membership inference and differential identifiability experiments that we will use for comparison between the DP adversary and the MI adversary throughout this work.

2.1 Differential Privacy

We define data analysis as the evaluation of a function $f : DOM \rightarrow \mathcal{R}$ on a dataset $\mathcal{D}$ from domain DOM yielding a result r from the set of all results $\mathcal{R}$. DOM is assumed to be a finite set, and D consists of independently sampled values from a distribution over DOM [LQS$^+$13]. Since r is computed from $\mathcal{D}$, r inevitably leaks information about the respective entries $d \in D$ (cf. impossibility of Dalenius' desideratum [Dwo06]). Differential Privacy [Dwo06] offers an anonymization guarantee for statistical functions such as those found in data analysis. In contrast to semantic anonymization (e.g., k-anonymity [SS98, Sam01]), DP perturbs the result of a function $f(\cdot)$ over a dataset $\mathcal{D} = \{d_1, \ldots, d_n\}$ s.t. the result of $f(\cdot)$ could have been produced from dataset $\mathcal{D}$ or some *neighboring* dataset $\mathcal{D}'$. A neighboring dataset $\mathcal{D}'$ either differs in the presence of one data point from $\mathcal{D}$ (unbounded DP) or in the value of one data point in $\mathcal{D}$ (bounded DP). Thus, plausible deniability is provided to participants in the dataset D since their impact on the query function $f(\cdot)$ becomes bounded. DP provides a strong guarantee: protection against an adversary with knowledge of all points in a data set except one. However, one assumption of DP is that data points are independent; correlation between data points, such as those that may be found in social graph applications, cannot yield the same DP guarantees [KM11]. To add differentially private noise to the result of some arbitrary function $f(\cdot)$, *mechanisms* $\mathcal{M}$ according to Definition 1 are used. In the context of this work, we will assume w.l.o.g. that $\mathcal{D} \setminus \mathcal{D}' \neq \{\}$; in other words, $\mathcal{D}$ contains one datapoint more than $\mathcal{D}'$ when unbounded DP is considered.

Definition 1 ((ε, δ)-Differential Privacy [DKM$^+$06])**.** *A mechanism $\mathcal{M}$ gives (ε, δ)-Differential Privacy if for all independently sampled $\mathcal{D}, \mathcal{D}' \subseteq DOM$, where DOM is a finite set, differing in at most one element, and all possible mechanism outputs $\mathcal{S}$*

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

We define ε-DP as $(\varepsilon, \delta = 0)$-DP and refer to the application of a mechanism $\mathcal{M}$ to a function $f(\cdot)$ as *output perturbation*. The Gaussian mechanism is the predominant DP mechanism in machine learning for perturbing the outcome of stochastic gradient descent, and adds noise independently sampled from a Gaussian distribution centered at zero. Prior work [DR14] has analyzed

the tails of the normal distributions and found that bounding the standard deviation as follows fulfills (ε, δ) DP.

$$\sigma > \Delta f_2 \sqrt{2\ln(1.25/\delta)}/\varepsilon \tag{2.1}$$

σ depends not only on the DP guarantee, but also on a scaling factor Δf. DP holds if mechanisms are scaled to the global sensitivity Δf of Definition 2, i.e., the maximum contribution of a record in the dataset to the outcome of $f(\cdot)$. For example, in the case of counting queries Δf is usually 1, while for the sum of all salaries in a company Δf might be very large (e.g., reflecting the CEO salary). The DP guarantee is tight for any data point having an influence of Δf. Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring data sets, the global ℓ_2-sensitivity of a function f is defined as $\Delta f_2 = max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2$.

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

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

Note that the absolute global sensitivity as in Definition 2 can also be defined relative to local sensitivity as $\Delta f = \max_{\mathcal{D}} LS_f(\mathcal{D})$. The impact of local sensitivity, compared to using global sensitivity, is that less noise is added when ε is held constant, and ε is decreased when the noise distribution is held constant. Local sensitivity is specified in Definition 3 [NRS07] and scales the differential privacy protection to a fixed dataset $\mathcal{D}$.

Definition 3 (Local Sensitivity)**.** *Let $\mathcal{D}$ and $\mathcal{D}'$ be neighboring. For a given finite set DOM, independently sampled dataset $\mathcal{D} \subseteq$ DOM, 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 the differentially private stochastic gradient descent, perturbed outputs are released repeatedly in an iterative process. The most basic form of composition for accounting of 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 ($\sum_i \varepsilon_i$, $\sum_i \delta_i$)-DP. However, sequential composition adds more noise than necessary. A tighter analysis of composition is provided by Mironov [Miv17]: $(\alpha, \varepsilon_{RDP})-$Rényi Differential Privacy (RDP) describes the difference in distributions $\mathcal{M}(\mathcal{D}), \mathcal{M}(\mathcal{D}')$ by their Rényi divergence [vEH10]. For a sequence of k mechanism executions each providing (α, $\varepsilon_{RDP,i}$)-RDP, the total privacy guarantee composes to (α, $\sum_i \varepsilon_{RDP,i}$)-RDP. The ($\alpha$, ε_{RDP})-RDP guarantee converts to $(\varepsilon_{RDP} - \frac{\ln \delta}{\alpha - 1}, \delta)$-DP. The Gaussian mechanism is calibrated to RDP using the relation $\varepsilon_{RDP} = \alpha \cdot \Delta f_2^2 / 2\sigma^2$.

2.2 Membership Inference

Membership inference is a threat model for quantifying identifiability in machine learning. MI attacks are used to quantify how accurate an adversary can identify members of training data. Black-box MI assumes access to a trained machine learning model [SSSS17, JE19], and white-box MI extends the assumption to oberservations on the training data [NSH18]. Yeom et al. [YGFJ18] formalize MI in the following generic adversarial experiment:

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

1. *Sample $z_{\mathcal{D}}$ uniformly from $\mathcal{D}$ and $z_{\texttt{Dist}}$ from* `Dist`
2. *Choose $b \leftarrow \{0,1\}$ uniformly at random*
3. *Let*
$$z = \begin{cases} z_{\mathcal{D}} & \text{if } b\text{=}1 \\ z_{\texttt{Dist}} & \text{if } b\text{=}0 \end{cases}$$
4. *$\mathcal{A}_{MI}$ outputs $b' = \mathcal{A}_{MI}(\vec{r}, z, \texttt{Dist}, n, \mathcal{M}) \in \{0,1\}$. If $b' = b$, $\mathcal{A}_{MI}$ succeeds and the output of the experiment is 1, it is 0 otherwise*

Within this deliverable we aim to evaluate white-box attacks consistent with DP guarantees (i.e., auxiliary side knowledge). We thus extended the black-box experiment from Yeom et al. [YGFJ18] to consider not only mechanism outputs, but also the mechanism and, implicitly, the privacy parameters (ε, δ) themselves. This addition solely provides $\mathcal{A}_{MI}$ with additional information, so any upper bounds for the probability to succeed in this white-box experiment will hold for the black-box experiment. The probability of success in the above experiment is bound by the DP privacy parameter ε [YGFJ18]; however, the bound is very loose in practice [JE19].

2.3 Differential Identifiability

Lee et al. [LC11, LC12] introduce Differential Identifiability as a strong threat model for inferring membership in the input dataset of a function based on perturbed output. Differential Identifiability assumes the adversary to calculate the likelihood of all possible input datasets, so called *possible worlds* in a set Ψ, given a mechanism output. Li et al. [LQS+13] show that the Differential Identifiability threat model maps to the worst case against which bounded Differential Privacy protects when $|\Psi| = 2$, since DP considers two neighboring datasets $\mathcal{D}$, $\mathcal{D}'$ by definition, and possible worlds each have the same number of records. The DI adversary $\mathcal{A}_{\texttt{DI}}$ knows both neighboring datasets and receives the multidimensional output $\vec{r}$ of the mechanism applied to one of these two datasets. The adversary's task is to guess which of the two datasets $\mathcal{D}'$ or $\mathcal{D}$ was chosen as input. The experiment is similar to membership inference, since the attacker must decide whether the dataset contains the member that differs between the known $\mathcal{D}'$ and $\mathcal{D}$ or not. To allow us comparisons, we reformulate the original idea as a cryptographic experiment:

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

1. *Set $\vec{r}_{\mathcal{D}} := \mathcal{M}(\mathcal{D})$ and $\vec{r}_{\mathcal{D}'} := \mathcal{M}(\mathcal{D}')$*
2. *Choose $b \leftarrow \{0,1\}$ uniformly at random*
3. *Let*
$$\vec{r} = \begin{cases} \vec{r}_{\mathcal{D}}, & \text{if } b = 1 \\ \vec{r}_{\mathcal{D}'}, & \text{if } b = 0 \end{cases}$$

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

Since this experiment precisely defines an adversary with access to background knowledge on $\mathcal{D}$ and $\mathcal{D}'$, $\mathcal{A}_{\mathtt{DI}}$ is an implementable instance of the DP adversary [DR16].

3. Identifiability bounds for the differential privacy adversary in Machine Learning

In this chapter we formulate two bounds on identifiability of individual training records when releasing a differentially private machine learning model. The bounds hold for differential privacy under multidimensional queries and composition. While membership inference is commonly used for quantifying identifiability in machine learning, we consider the stronger DP adversary $\mathcal{A}_{\mathtt{DI}}$. We prove that protection against $\mathcal{A}_{\mathtt{DI}}$ also protects against $\mathcal{A}_{\mathtt{MI}}$ (Section 3.1). Afterwards, we introduce our identifiability bounds for $\mathcal{A}_{\mathtt{DI}}$ (Section 3.2). First, we define the identifiability risk as *adaptive posterior belief*, which is a new privacy metric for iterative mechanisms. Second, we discuss *membership advantage* for $\mathcal{A}_{\mathtt{DI}}$, which is a privacy metric complementing the adaptive posterior belief by offering a scaled quantification of the adversary's probability of success (i.e., how often posterior belief is greater than 50%).

3.1 Identifiability under the Strong Probabilistic Adversary

Societal norms such as identifiability w.r.t. anonymization are not matching to the mathematical concept behind DP, since DP limits the contribution of an individual to aggregated information [NW18]. For example, the European General Data Protection Regulation (GDPR) Recital 26 states that identifiability is determined by all means reasonably likely to be used to single out an individual [Eur16] and the American Health Insurance Portability and Accountability Act (HIPAA) explicitly requires identifiability guarantees in form of group sizes (i.e., 1/group size) in § 164.514 (2) [Ame10]. Rather, in a move from DP guarantees to societal norms, privacy parameter ε should be transformed to the probability of identifiability [CT13, CK12].

We propose the strong probabilistic adversary $\mathcal{A}_{\mathtt{DI}}$ as an alternative to using the MI adversary for evaluating DP guarantees [HMDD19, BGRK19, CYZF20, JE19, JWK+20, RRLM18]. DP aims to protect against adversaries with arbitrary auxiliary information, so intuitively, MI bounds based on DP guarantees will never be reached. The results of Jayaraman et al. [JE19] confirm this expectation empirically, citing a large gap between the theoretical membership advantage bound formulated by DP ($e^{\varepsilon} - 1$) and the empirical membership advantage, which implies that more powerful inference attacks exist. $\mathcal{A}_{\mathtt{DI}}$ performs such a stronger attack, which can also be implemented and yields metrics related to identifiability. $\mathcal{A}_{\mathtt{DI}}$ has access to arbitrary auxiliary information as is assumed in the original DP guarantee. $\mathcal{A}_{\mathtt{DI}}$ also operates in a white-box model and consequently observes all training steps of a machine learning algorithm, a characteristic especially found in federated learning. Therefore, $\mathcal{A}_{\mathtt{DI}}$ quantifies what the strongest possible DP adversary can infer. We first show that protection against DI implies protection against MI. Equivalently, we show that $\mathcal{A}_{\mathtt{DI}}$ is stronger than $\mathcal{A}_{MI}$ due to the additional available auxiliary information. Concretely, the $\mathcal{A}_{\mathtt{DI}}$ knows both neighboring data sets $\mathcal{D}$ and $\mathcal{D}'$ instead of only receiving one value z and the size n of the data set from which the data points are drawn.

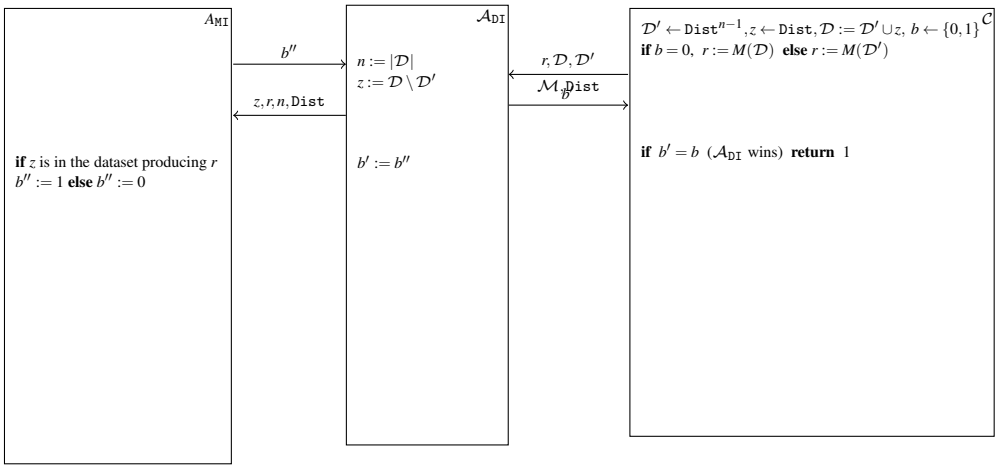


Figure 3.1: Reduction of $\mathcal{A}_{\mathtt{DI}}$ to $\mathcal{A}_{MI}$, shown here for unbounded DP

Proposition 1. *Differential identifiability implies membership inference: if $\mathcal{A}_{MI}$ wins Exp_{MI}, then one can construct $\mathcal{A}_{\mathtt{DI}}$ that wins Exp_{DI}, as shown in Figure 3.1.*

Proof. We prove the proposition by contradiction: assume that the mechanism $\mathcal{M}$ successfully protects against $\mathcal{A}_{\mathtt{DI}}$, but that there exists an adversary $\mathcal{A}_{MI}$ that wins Exp_{MI}. Again, we assume w.l.o.g. that $\mathcal{D} \setminus \mathcal{D}' \neq \{\}$; otherwise, we can formulate an analogous proof. We construct an adversary $\mathcal{A}_{\mathtt{DI}}$ that also wins Exp_{DI} as follows, see also Figure 3.1:

1. On inputs $\mathcal{D}, \mathcal{D}', \mathcal{M}, \vec{r}, \mathtt{Dist}$, $\mathcal{A}_{\mathtt{DI}}$ calculates $n = |\mathcal{D}|$ and let $z = \mathcal{D} \setminus \mathcal{D}'$.
2. $\mathcal{A}_{\mathtt{DI}}$ gives $(z, \vec{r}, n, \mathtt{Dist})$ to $\mathcal{A}_{MI}$.
3. $\mathcal{A}_{MI}$ gives $b'' = \mathcal{A}_{MI}(z, \vec{r}, n, \mathtt{Dist})$ to $\mathcal{A}_{\mathtt{DI}}$ in response.
4. $\mathcal{A}_{\mathtt{DI}}$ outputs b'.

By the definition of Exp_{DI}, $\mathcal{A}_{\mathtt{DI}}$ wins if $b' = b$, and thus succeeds in the following cases.
Case 1: $b = 1$, which means $\vec{r} = \mathcal{M}(\mathcal{D})$. Since $z \in \mathcal{D}$, this is exactly the case where $\mathcal{A}_{MI}$ correctly outputs $b' = 1$. Therefore $b' = b$.
Case 2: $b = 0$, which means $\vec{r} = \mathcal{M}(\mathcal{D}')$. Since $z \notin \mathcal{D}'$, this is exactly the case where $\mathcal{A}_{MI}$ correctly outputs $b' = 0$. Therefore $b' = b$. For both cases $\mathcal{A}_{\mathtt{DI}}$ wins ($b' = b$), which contradicts the assumption that the mechanism $\mathcal{M}$ successfully protects against $\mathcal{A}_{\mathtt{DI}}$. It is more difficult for a mechanism to protect against Exp_{DI} than against Exp_{MI}, which is equivalent to the statement that if $\mathcal{A}_{MI}$ wins Exp_{MI}, then $\mathcal{A}_{\mathtt{DI}}$ wins Exp_{DI} as well. □

3.2 Posterior Belief and Advantage

We will now introduce two identifiability metrics for the differentially private gradient descent under $\mathcal{A}_{\mathtt{DI}}$. In line with Li et al. [LQS+13] we also assume that datasets $\mathcal{D}$ and $\mathcal{D}'$ to be sampled mutually independently from the identical distribution (i.i.d.) `Dist` over datasets. Under this assumption the samples of the dataset are essentially sampled independently, however, potentially with different probabilities (i.e., stratified).

Under this assumption the samples of the dataset are essentially sampled independently, however, potentially with different probabilities. Note that this assumption includes not only the usual i.i.d. assumption, but also stratified sampling.

Here we model these algorithms as iterative algorithms where each unperturbed gradient is modeled as a function evaluation of the form $\vec{r}_{t+1} = f(\mathcal{D}, \vec{\tilde{r}}_t) = f_t(\mathcal{D})$. In the case of gradient descent we update the weights $\vec{r}$ using the perturbed gradients $\tilde{g}_t = \mathcal{M}_t(\mathcal{D}, \vec{r}_{t-1})$ and learning rate α as $\vec{\tilde{r}}_{t+1} = \vec{\tilde{r}}_t + \alpha \cdot \tilde{g}_t$. The initial value can be modeled as $\vec{r}_0$, and the functions and the mechanism $\mathcal{M}$ may thus change during the iterations. This iterative procedure requires an extension of $\mathcal{A}_{\mathtt{DI}}$ to composition, where $\mathcal{A}_{\mathtt{DI}}$ receives a matrix $R = \vec{r}_0, \ldots, \vec{r}_T$ that consists of the results from all training steps. In this deliverable we will analyze these steps under DP composition.

We quantify individual identifiability by the Bayesian posterior belief in the right dataset $\mathcal{D}$ compared to the belief in $\mathcal{D}'$. Deriving a tight upper bound for the posterior belief of $\mathcal{A}_{\mathtt{DI}}$ in Section 4.1 then results in the highest chance of successful inference of an individual.

Definition 4 (Adaptive Posterior Belief)**.** *Consider the setting of Experiment 2 and denote $R_i = (\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_i)$ as the result matrix, comprising i multidimensional mechanism results. The posterior belief in the correct dataset $\mathcal{D}$ is defined as the probability conditional 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.

While the posterior belief is in principle a sophisticated probability distribution, it is now shown that the iterative procedure leads to a significant simplification: each β_i can be computed from the previous β_{i-1}. It turns out, that the final belief can be computed using the following Lemma 1 which we will use later to further analyze the strongest possible attacker $\mathcal{A}_{\mathtt{DI}}$ of Experiment 2.

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

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

Proof. We prove the lemma by iteration over k.
$k = 1$: We assume the attacker starts with uniform priors $\Pr(\mathcal{D}) = \Pr(\mathcal{D}') = \frac{1}{2}$. Thus, $\beta_1(\mathcal{D}|R_1)$ can be directly calculated using the definition and division of both numerator and denominator by the numerator:

$$\begin{aligned}\beta_1(\mathcal{D}|R_1) &= \frac{\Pr(\mathcal{M}_1(\mathcal{D}) = \vec{r}_1)}{\Pr(\mathcal{M}_1(\mathcal{D}) = \vec{r}_1) + \Pr(\mathcal{M}_1(\mathcal{D}') = \vec{r}_1)} \\ &= \frac{1}{1 + \frac{\Pr(\mathcal{M}_1(\mathcal{D}') = \vec{r}_1)}{\Pr(\mathcal{M}_1(\mathcal{D}) = \vec{r}_1)}}\end{aligned}$$

$k-1 \to k$: In the second step $\beta_{k-1}(\mathcal{D}|R_{k-1})$ is used as the prior, using the shorthand notations $\beta_k := \beta_k(\mathcal{D}|R_k)$, and in the last step $p_k := \Pr(\mathcal{M}_k(\mathcal{D}) = \vec{r}_k)$ and $p'_k := \Pr(\mathcal{M}_k(\mathcal{D}') = \vec{r}_k)$ the calculation of $\beta_k(\mathcal{D}|R_k)$ starts as for the induction start $k = 1$

$$\begin{aligned}\beta_k &= \frac{\Pr(\mathcal{M}_k(\mathcal{D}) = \vec{r}_k) \cdot \beta_{k-1}}{\Pr(\mathcal{M}_k(\mathcal{D}) = \vec{r}_k) \cdot \beta_{k-1} + \Pr(\mathcal{M}_k(\mathcal{D}') = \vec{r}_k) \cdot (1-\beta_{k-1})} \\ &= \frac{1}{1 + \frac{\Pr(\mathcal{M}_k(\mathcal{D}')=\vec{r}_k) - \Pr(\mathcal{M}_k(\mathcal{D}')=\vec{r}_k)\cdot\beta_{k-1}}{\Pr(\mathcal{M}_k(\mathcal{D})=\vec{r}_k)\cdot\beta_{k-1}}} \\ &= \frac{1}{1 + \frac{p'_k - p'_k\beta_{k-1}}{p_k\beta_{k-1}}}\end{aligned}$$

Now the induction assumption can be substituted for the right term of the denominator and then multiplying the numerator and denominator with $\prod_{i=1}^{k-1} p_i + \prod_{i=1}^{k-1} p'_i$ leads to

$$\begin{aligned}\frac{p'_k - p'_k\beta_{k-1}}{p_k\beta_{k-1}} &= \frac{p'_k - p'_k \frac{\prod_{i=1}^{k-1} p_i}{\prod_{i=1}^{k-1} p_i + \prod_{i=1}^{k-1} p'_i}}{p_k \frac{\prod_{i=1}^{k-1} p_i}{\prod_{i=1}^{k-1} p_i + \prod_{i=1}^{k-1} p'_i}} \\ &= \frac{p'_k(\prod_{i=1}^{k-1} p_i + \prod_{i=1}^{k-1} p'_i) - p'_k \prod_{i=1}^{k-1} p_i}{p_k \prod_{i=1}^{k-1} p_i} \\ &= \frac{\prod_{i=1}^{k} p'_i}{\prod_{i=1}^{k} p_i}\end{aligned}$$

where in the last step the first and the third term in the denominator cancel out and lead to the desired result when inserted back into the last form of β_k above. □

The above proof illustrates that $\mathcal{A}_{\mathtt{DI}}$ behaves as a binary classifier that chooses the option w.r.t. to the highest posterior probability. Specifically, $\mathcal{A}_{\mathtt{DI}}$ computes posterior beliefs $\beta(\cdot)$ for datasets $\mathcal{D}$ and $\mathcal{D}'$ and guesses the dataset

$$\underset{D \in \{\mathcal{D}, \mathcal{D}'\}}{\arg\max}\, \beta(D|R_k).$$

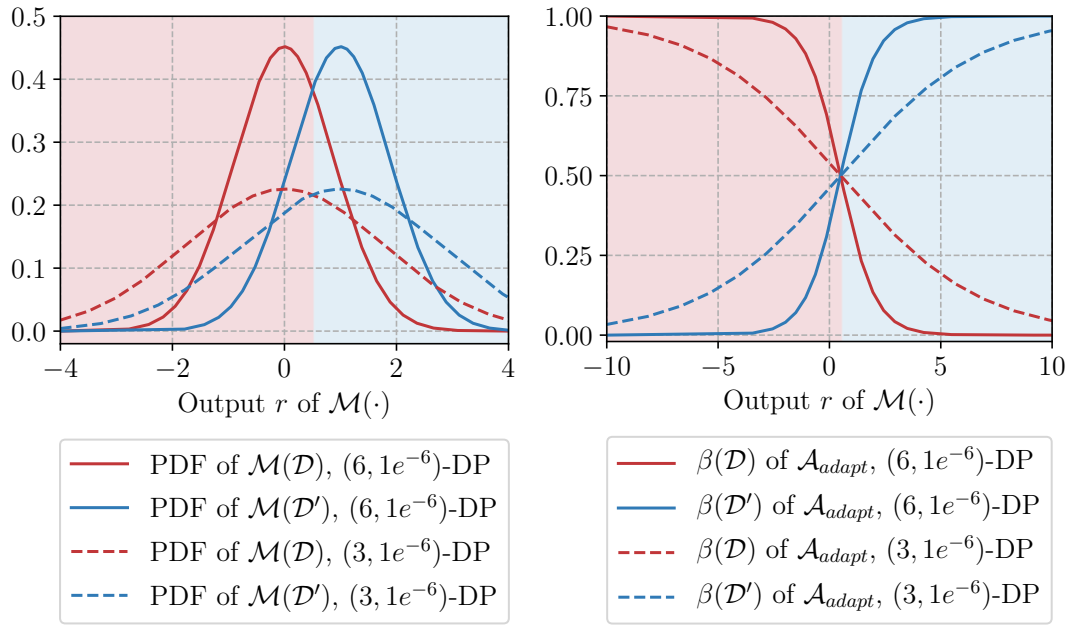
This decision process can be simplified as follows: the probabilistic mechanism $\mathcal{M}$ turns R_k into random variables which are denoted as

$$X_1 := \mathcal{M}(\mathcal{D}) \text{ and } X_0 := \mathcal{M}(\mathcal{D}') \tag{3.1}$$

corresponding to the cases $b = 1$ and $b = 0$, respectively. Since $\mathcal{A}_{\mathtt{DI}}$ knows $\mathcal{D}, \mathcal{D}'$ and the mechanism $\mathcal{M}$, $\mathcal{A}_{\mathtt{DI}}$ also knows the corresponding probability densities g_{X_1} and g_{X_0}. The densities have the same shape depending on the mechanism but are centered at the different unperturbed results $f(\mathcal{D})$ and $f(\mathcal{D}')$, respectively, as visualized in Figure 3.2(a) with $f(\mathcal{D}) = 0, f(\mathcal{D}') = 1$ for multiple DP guarantees. Assuming uniform prior beliefs, $\mathcal{A}_{\mathtt{DI}}$'s decision then depends on whether $\mathcal{A}_{\mathtt{DI}}$ believes that R_k stems more likely from X_1 or X_0 and therefore decides

$$\mathcal{A}_{\mathtt{DI}}(R_k, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist}) = \underset{b \in \{0,1\}}{\arg\max}\, g_{X_b}(R_k) \tag{3.2}$$

More generically, if we choose not to assume uniform prior beliefs, $\mathcal{A}_{\mathtt{DI}}$ instead chooses the dataset that results in a larger posterior belief. The posterior belief in our simple example is visualized in Figure 3.2(b). So $\mathcal{A}_{\mathtt{DI}}$ is essentially a naive Bayes classifier whose decision boundary



(a) Decision boundary based on PDFs

(b) Decision boundary based on confidence

Figure 3.2: The decision boundary of $\mathcal{A}_{\texttt{DI}}$ does not change when increasing the privacy guarantee since (ε, δ) causes the PDFs of $\mathcal{D}$ and $\mathcal{D}'$ to become squeezed. Thus $\mathcal{A}_{\texttt{DI}}$ will exclusively choose $\mathcal{D}$ if a value is sampled from the left, red region, and vice versa for $\mathcal{D}'$ in the right, blue region. Still, confidence towards either decision declines.

is depicted by the change in background color in Figure 3.2(b). The input features are the perturbed results R_k, and the exact probability distributions of each class is known. Note that the distributions are entirely defined by $\mathcal{D}$, $\mathcal{D}'$, and the mechanism $\mathcal{M}$, so $\mathcal{A}_{\texttt{DI}}$ does not use the knowledge of the distribution `Dist` from which $\mathcal{D}$ and $\mathcal{D}'$ were sampled.

The posterior belief quantifies the probability for the original dataset $\mathcal{D}$ for a single optimization procedure resulting in the specific results R_k; however, in another optimization instance the result R_k could differ. In Section 4.1, we will therefore define an upper bound on $\beta(\mathcal{D})$.

We expect the question "How likely is it that an adversary correctly guessed the data set in which I have participated?" to be a major point of interest when interpreting DP guarantees in iterative evaluations of $\mathcal{M}$, like those found in data science use cases such as machine learning. The bound ρ_β for the posterior belief indicates the worst case probability of identifying that a given, single record belongs to the dataset $\mathcal{D}$. When posterior belief ρ_β is low, an individual can plausibly deny that the hypothesis of $\mathcal{A}_{\texttt{DI}}$ is correct. In practice, it may be even more important to know how often $\mathcal{A}_{\texttt{DI}}$ makes a correct hypothesis over the course of many runs, which only occurs when $\rho_\beta > 50\%$. This is quantified by the *advantage*, which is the success rate normalized to the range $[-1, 1]$, where 0 corresponds to random guessing.

Definition 5 (Advantage)**.** *Consider an experiment Exp and denote $R_i = (\vec{r}_0, \vec{r}_1, \ldots, \vec{r}_i)$ as the result matrix, comprising i multidimensional mechanism results. The advantage is then defined as*

$$Adv{=}2\Pr(Exp = 1) - 1 \tag{3.3}$$

where the probability is over the random iterative choices of the mechanisms up to step k. When we consider Exp^{DI}, we define the corresponding advantage as Adv^{DI}. When we consider Exp^{MI},

we define the corresponding advantage as Adv^{MI}.

In Section 4.2 tight upper bounds for the advantage will be derived. The advantage and its upper bounds may be important ingredients for the technical implementation of legal privacy requirements that formulate anonymization breaches in terms of individual identifiability.

4. Derivation of tight upper bounds

Within this Chapter we will derive tight upper bounds for the previously introduced identifiability metrics *posterior belief* and *advantage* in Section 4.1 and Section 4.2. Afterwards, we will illustrate how the bounds can be transformed into privacy parameters under RDP composition in Section 4.3.

4.1 Posterior Belief Bound

Bayesian posterior belief was introduced as a measurement of identifiability. We formulate a generic bound on posterior belief independent of datasets $\mathcal{D}$ and $\mathcal{D}'$, the mechanism $\mathcal{M}$, and the mechanism output matrix R, which consists of multiple multidimensional mechanism outputs $\vec{r}_i$. The proposed bound solely assumes that the DP bound holds and makes no further simplifications or assumptions, which results in an identifiability-based interpretation of DP guarantees.

Theorem 2 shows that $\mathcal{A}_{\mathtt{DI}}$ operates under the sequential composition theorem. The sequential composition theorem states that given $\mathcal{M}_i$ providing ε_i-Differential Privacy, the sequence of $\mathcal{M}_{1,\dots,k}(\mathcal{D})$ provides $\sum_{1,\dots,k}\varepsilon_i$-DP [DR14]. In the following theorem this is not only shown for ε-Differential Privacy but also for (ε,δ)-Differential Privacy setting which is common when using the differentially private stochastic gradient descent in the machine learning setting [ACG+16].

Theorem 2 (Bounds for the Adaptive Posterior Belief)**.** *Consider experiment Exp^{DI} with neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$.*
(i) Let $\mathcal{M}_1,\dots,\mathcal{M}_k$ be a sequence of arbitrary but independent differentially private learning algorithms providing $\varepsilon_1,\dots,\varepsilon_k$-Differential Privacy to functions f_i with multidimensional output. Then the strong probabilistic adversary's posterior belief is bounded by

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

(ii) Let $\mathcal{M}_1,\dots,\mathcal{M}_k$ be a sequence of arbitrary but independent differentially private learning algorithms where each $\mathcal{M}_i$ provides (ε_i,δ_i)-Differential Privacy to multidimensional functions f_i. Then the same bound as above holds with probability $1-\sum_{i=1}^{k}\delta_i$.

Proof. (i) The adversary with unbiased prior (i.e., 0.5) regarding neighboring datasets $\mathcal{D},\mathcal{D}'$ has a maximum posterior belief of $1/(1+e^{-\varepsilon})$ when the ε-differentially private Laplace mechanism is applied to a function with a scalar output [LC12]. This upper bound holds also for arbitrary ε-differentially private learning algorithms with multidimensional output. We bound the general belief calculation by the inequality of Definition 1. According to DP, for any differentially private mechanism result $\vec{r}$, where differentially private mechanism $\mathcal{M}(\mathcal{D})$ has been trained with dataset $\mathcal{D}$ and $\mathcal{M}(\mathcal{D}')$ has been trained with dataset $\mathcal{D}'$:

$$\Pr(\mathcal{M}(\mathcal{D}) = \vec{r}) \leq e^{\varepsilon} \Pr(\mathcal{M}(\mathcal{D}') = \vec{r}) + \delta$$

Assuming equal priors, the posterior belief can be calculated as follows:

$$\begin{aligned}\beta(\mathcal{D}|R) &= \frac{1}{1 + \frac{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)}{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D})=\vec{r}_i)}} \\ &\leq \frac{1}{1 + \frac{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)}{\prod_{i=1}^{k} e^{\varepsilon_i} \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)+\delta_i}}\end{aligned}$$

For $\delta = 0$, the last equation simplifies to:

$$\begin{aligned}\beta(\mathcal{D}|R) &\leq \frac{1}{1 + \frac{\prod_{i=1}^{k} \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)}{\prod_{i=1}^{k} e^{\varepsilon_i} \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)}} \\ &= \frac{1}{1 + \prod_{i=1}^{k} e^{-\varepsilon_i}} = \frac{1}{1 + e^{-\sum_{i=1}^{k} \varepsilon_i}} = \rho_\beta\end{aligned}$$

□

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

$$\varepsilon = ln\left(\frac{\rho_\beta}{1-\rho_\beta}\right)$$

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.

4.2 Bound for Expected Membership Advantage

Now we derive the upper bound for the advantage of $\mathcal{A}_{\mathtt{DI}}$ (Adv^{DI}). Membership advantage is used to quantify the success of membership inference adversaries, and bounds in terms of DP ε have been previously derived for MI [YGFJ18]. First we derive that this general bound for $\mathcal{A}_{MI}$ also holds for $\mathcal{A}_{\mathtt{DI}}$, as expected based on Proposition 1. We then derive a tighter bound for the (ε, δ)-differentially private Gaussian mechanism, which is commonly used for machine learning, by modeling $\mathcal{A}_{\mathtt{DI}}$ as a naive Bayes classifier.

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

$$\mathrm{Adv}^{\mathrm{DI}} \leq (e^{\varepsilon} - 1) \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0) \tag{4.1}$$

Proof. First the definition is re-written using both ways to success in the experiment. Then using that both datasets are chosen equally likely by the challenger ($\Pr(b = 1) = \Pr(b = 0) = 1/2$), substituting $\Pr(b' = 0 \,|b = 0)$ by the probability of the complementary event $1 - \Pr(b' = 1 \,|b = 0)$) and finally substituting $b' = 1$ by $\mathcal{A}_{\mathtt{DI}} = 1$ leads to the formula (4.2) of Yeom et al. [YGFJ18]

$$\begin{aligned}\mathrm{Adv}^{\mathrm{DI}} &= 2(\Pr(b = 1) \Pr(b' = 1 \,|b = 1) + \\ &\quad + \Pr(b = 0) \Pr(b' = 0 \,|b = 0)) - 1 \\ &= \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 1) - \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0)\end{aligned} \tag{4.2}$$

which is the difference between the probability for true detection of $\mathcal{D}$ minus the probability of incorrectly choosing $\mathcal{D}$. Now we use the fact that the mechanism $\mathcal{M}$ turns r into random variables $X_1 := \mathcal{M}(\mathcal{D})$ and $X_0 := \mathcal{M}(\mathcal{D}')$ for the cases $b = 1$ and $b = 0$, respectively. We formulate the probability density functions as g_{X_1} and g_{X_0}. Additionally $A(r)$ is introduced as a shorthand for $\mathcal{A}_{\mathtt{DI}}(\vec{r}, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist})$

$$\begin{aligned}
\mathrm{Adv}^{\mathrm{DI}} &= \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | r = \mathcal{M}(\mathcal{D})) - \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | r = \mathcal{M}(\mathcal{D}')) \\
&= \mathbb{E}_{r=\mathcal{M}(\mathcal{D})}(\mathcal{A}_{\mathtt{DI}}(\vec{r}, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist})) - \\
&\quad \mathbb{E}_{r=\mathcal{M}(\mathcal{D}')}(\mathcal{A}_{\mathtt{DI}}(\vec{r}, \mathcal{D}, \mathcal{D}', \mathcal{M}, \mathtt{Dist})) \\
&= \int g_{X_1}(\vec{r}) A(\vec{r}) \mathrm{d}\vec{r} - \int g_{X_0}(\vec{r}) A(\vec{r}) \mathrm{d}\vec{r} \qquad (4.3) \\
&= \int (g_{X_1}(\vec{r}) - g_{X_0}(\vec{r})) A(\vec{r}) \mathrm{d}\vec{r} \qquad (4.4)
\end{aligned}$$

Since ε-DP is defined as $Pr(\mathcal{M}(\mathcal{D}) \in S) \leq e^{\varepsilon} Pr(\mathcal{M}(\mathcal{D}') \in S)$, it yields to the same inequality $g_{X_1} \leq e^{\varepsilon} g_{X_0}$ for the densities for all S (i.e., at each point),

$$\begin{aligned}
\mathrm{Adv}^{\mathrm{DI}} &\leq (e^{\varepsilon} - 1) \int g_{X_0}(\vec{r}) A(r) \mathrm{d}\vec{r} \\
&= (e^{\varepsilon} - 1) \Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0) \qquad (4.5) \\
&\leq e^{\varepsilon} - 1
\end{aligned}$$

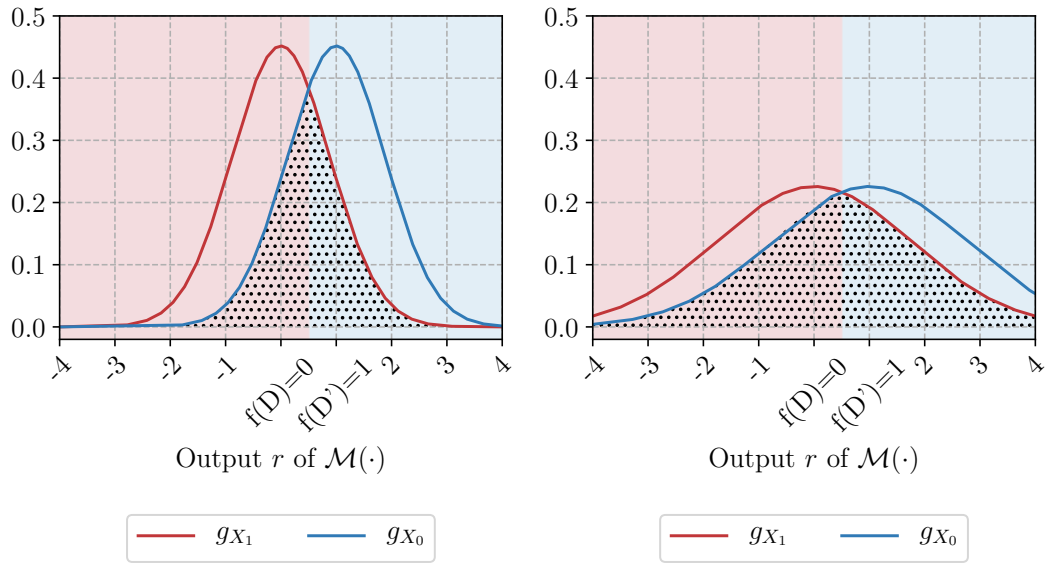
□

Bounding $\Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0)$ by 1 results in $\mathrm{Adv}^{\mathrm{DI}} \leq e^{\varepsilon} - 1$. Since the mechanism preserves some utility, a rational adversary $\mathcal{A}_{\mathtt{DI}}$ that tries to win will make a correct guess at least 50% of the time, so $\Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0) \leq 0.5$. Substituting this into Equation (4.1) yields to a tighter bound $\mathrm{Adv}^{\mathrm{DI}} \leq (e^{\varepsilon} - 1)/2$.

When $\mathcal{A}_{\mathtt{DI}}$ acts as a naive Bayes classifier, only a complete lack of utility from infinite noise results in $\Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0) = 0.5$. Otherwise, $\Pr(\mathcal{A}_{\mathtt{DI}} = 1 | b = 0)$ is far smaller than 0.5; therefore, even this membership advantage bound is usually not tight. Since protection against DI implies protection against MI, as proven in Proposition 1, the bound also holds for $\mathcal{A}_{MI}$. A similar bound for advantage has been proven for the MI adversary [YGFJ18], $e^{\varepsilon} - 1$; however, $(e^{\varepsilon} - 1)/2$ is smaller and will therefore also be tighter. $\mathcal{A}_{\mathtt{DI}}$ should also come closer than $\mathcal{A}_{MI}$ to the bound. This is in line with Jayaraman et al. [JE19] who expect that this would be the case for a stronger inference attack than MI.

We now derive a tighter bound ρ_a on Adv^{DI} for the Gaussian mechanism and (ε, δ)-differential privacy, continuing from Equation (4.4). Note that under the assumption of equal priors, the strongest possible adversary of Equation (3.2) maximizes (4.4) by choosing $b = 1$ if $(g_{X_1}(\vec{r}) - g_{X_0}(\vec{r})) > 0$ and $b = 0$ otherwise. The resulting bound on Adv^{DI} is constructed from $\mathcal{A}_{\mathtt{DI}}$'s strategy; however, it holds for all weaker adversaries, including $\mathcal{A}_{MI}$. Since we argue that $\mathcal{A}_{\mathtt{DI}}$ 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.

Now, since $\mathcal{A}_{\mathtt{DI}}$ is essentially a naive Bayes classifier with known probability distributions, we can use the properties of normal distributions (we refer to Tumer et al. [TG96] for full details). Looking at the decision boundary of this classifier (i.e., when to choose $\mathcal{D}$ or $\mathcal{D}'$) under $\mathcal{M}_{Gau}$ with different (ε, δ) guarantees, we find that the decision boundary does not change as long as the



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

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

Figure 4.1: For visualization purposes, we arbitrarily set $f(\mathcal{D}) = 0, f(\mathcal{D}') = 1$. The plots show $\mathcal{A}_{\mathtt{DI}}$ error regions for varying ε, $\mathcal{M}_{Gau}$, $f(\mathcal{D}), f(\mathcal{D}')$. Note that the probability density functions and thus the regions under the curve are not scaled by the prior.

probability density functions (PDF) are symmetric. For example, consider the given datasets $\mathcal{D}$, $\mathcal{D}'$ and mechanism $\mathcal{M} : DOM \to \mathbb{R}$ that yield $f(\mathcal{D}) = 0$ and $f(\mathcal{D}') = 1$ without noise. Furthermore, assuming w.l.o.g. that $\Delta f_2 = 1$. If a $(6, 10^{-6})$-DP $\mathcal{M}_{Gau}$ is applied for perturbation, $\mathcal{A}_{\mathtt{DI}}$ has to choose between the two PDFs in Figure 4.1(a) based on the output $\mathcal{M}(\cdot) = r$. The regions where $\mathcal{A}_{\mathtt{DI}}$ chooses $\mathcal{D}$ are shaded red in both figures, and regions that result in the choice $\mathcal{D}'$ are shaded blue. Increasing the privacy guarantee to $(3, 10^{-6})$-DP in Figure 4.1(b) squeezes the PDFs and confidence curves. However, the decision boundary of the regions at which $\mathcal{A}_{\mathtt{DI}}$ chooses a certain dataset stay the same. Thus, it is important to note that holding r constant and reducing (ε, δ) solely affects the posterior belief of $\mathcal{A}_{\mathtt{DI}}$, not the choice (i.e., the order from most to least confident is maintained even while maximum posterior belief is lowered). The corresponding regions of error are shaded in Figures 4.1(a) and 4.1(b), where we see that a stronger guarantee reduces Adv^{DI}.

We assumed throughout our work that $\mathcal{A}_{\mathtt{DI}}$ has uniform prior beliefs on the possible databases $\mathcal{D}$ and $\mathcal{D}'$ over which $f()$ was evaluated. This distribution is iteratively updated based on the posterior resulting from the mechanism output r. Thus, $\mathcal{A}_{\mathtt{DI}}$ is essentially representing a Bayesian classifier. This allows us to analyze the concrete distributions resulting from the upper bounds ρ_β and ρ_a and the mechanism $\mathcal{M}$. If, for example, $\mathcal{M}_{Gau}$ is applied to achieve (ε, δ)-DP, we can determine the expected membership advantage of the practical attacker $\mathcal{A}_{\mathtt{DI}}$ analytically by the overlap of the resulting Gaussian distributions [MKB79, 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) $\vec{\mu}_1 = f(\mathcal{D}), \vec{\mu}_2 = f(\mathcal{D}')$.

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 prob-*

abilistic adversary on either data set $\mathcal{D}, \mathcal{D}'$.

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

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

Proof. The starting point for the derivation is Equation (4.3) where the Gauss-distributions are to be inserted for g_{X_1} and g_{X_0}. Since both distributions arise from the same mechanism, they have the same Σ but different means $\mu_1 = f(\mathcal{D})$ and $\mu_0 = f(\mathcal{D}')$. Since the strongest adversary is the Bayes adversary that chooses according to Equation (3.2), and we assume equal priors, the decision boundary between $\mathcal{D}$ and $\mathcal{D}'$ is the point of intersection of the densities (see Figure 4.1(a) for the 1D-case). In general, this is exactly the situation of linear discriminant analysis where it is known to be a hyperplane halfway between $\mu_1 = f(\mathcal{D})$ and $\mu_0 = f(\mathcal{D}')$. Mathematically, by setting $\ln(g_{X_1}) = \ln(g_{X_1})$, the plane can be calculated to be halfway $(\Delta/2)$ between the two centers, where Δ is the Mahalanobis distance [Mah36] $\Delta := \sqrt{(\vec{\mu}_1 - \vec{\mu}_2)^T \Sigma^{-1} (\vec{\mu}_1 - \vec{\mu}_2)}$. Notably the decision boundary between $\mathcal{D}$ and $\mathcal{D}'$ notably does not depend on Σ and therefore ε, but the distance between $f(\mathcal{D})$ and $f(\mathcal{D}')$ (i.e., sensitivity). As we add independent noise in all dimensions $\Sigma = \sigma^2 \mathbb{I}$, integration in Equation (4.3) normal to this direction leads to factors 1 and only the 1-D integration along the direction through the 2 centers remains with Δ simplified to $\frac{\|\vec{\mu}_1 - \vec{\mu}_2\|_2}{\sigma}$. Thus,

$$\begin{aligned}
\text{Adv}^{\text{DI}} &= \Phi(\Delta/2) - \Phi(-\Delta/2) \\
&= 2\Phi(\Delta/2) - 1 \\
&= 2\Phi\left(\frac{\|\vec{\mu}_1 - \vec{\mu}_2\|_2}{2\sigma_i}\right) - 1
\end{aligned}$$

Inserting the standard deviation needed for (ε, δ)-DP from (2.1) then yields

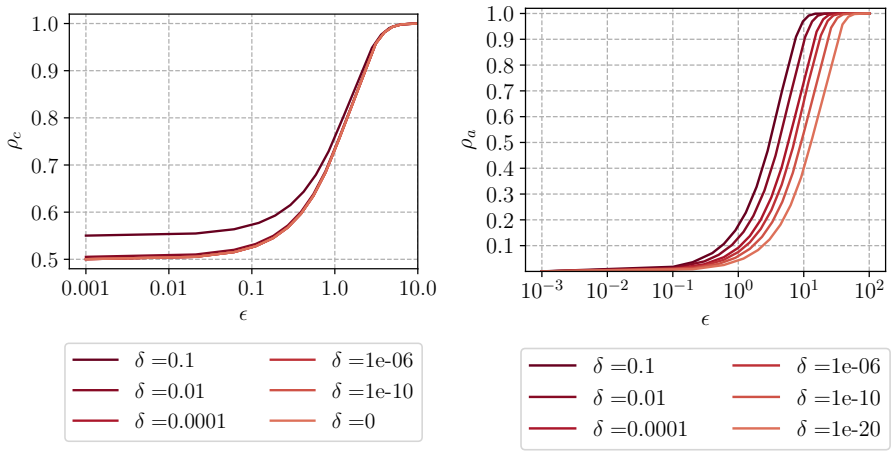
$$\begin{aligned}
\text{Adv}^{\text{DI}} &= 2\Phi\left(\frac{\|\vec{\mu}_1 - \vec{\mu}_2\|_2}{2\Delta f_2(\sqrt{2\ln(1.25/\delta)}/\varepsilon)}\right) - 1 \\
&\leq 2\Phi\left(\frac{\varepsilon}{2(\sqrt{2\ln(1.25/\delta)})}\right) - 1 = \rho_a
\end{aligned}$$

□

The theorem can be used to calculate ε from a chosen maximum expected advantage

$$\varepsilon = \sqrt{2\ln(1.25/\delta)}\,\Phi^{-1}\left(\frac{\rho_a + 1}{2}\right)$$

As for the posterior belief, the (ε, δ) guarantees with $\delta > 0$ can be expressed via a scalar value ρ_a. However, a specific membership advantage must be computed individually for different kinds of mechanisms $\mathcal{M}$. We provide plots of ρ_β and ρ_a for different (ε, δ) in Figure 4.2. For ρ_a, the curves are specific for $\mathcal{M}_{Gau}$. In contrast, ρ_β is independent of $\mathcal{M}$. To compute both measures, we use Theorem 2 and Theorem 3. We also assume w.l.o.g. that $f(\mathcal{D}) = (0_1, 0_2, \ldots, 0_k)$ and $f(\mathcal{D}') = (1_1, 1_2, \ldots, 1_k)$ for all dimensions k. Thus, $f(\mathcal{D})$ and $f(\mathcal{D}')$ are maximally distinguishable, resulting in $\Delta f_2 = \sqrt{k}$. Figure 4.2(a) illustrates that there is no significant difference between the adaptive posterior belief ρ_β for ε-DP and (ε, δ)-DP for $0 < \delta < 0.1$. In contrast, ρ_a strongly depends on the choice of δ as depicted in Figure 4.2(b). For example, ρ_a is low for $(2, 10^{-6})$-DP indicating that the probability of $\mathcal{A}_{\text{DI}}$ choosing $\mathcal{D}$ is similar to choosing $\mathcal{D}'$. Yet, the corresponding ρ_β is high, which provides support that $\mathcal{A}_{\text{DI}}$ guesses is correct.



(a) ρ_β (b) ρ_a

Figure 4.2: The expected adversarial worst-case confidence bound ρ_β and the adversarial membership advantage ρ_a for various (ε, δ) when using $\mathcal{M}_{Gau}$ for perturbation.

4.3 Bounds under composition with Renyi Differential Privacy

In practice, data owners can choose an overall privacy guarantee (ε, δ) according to values for the bounds ρ_β and ρ_α. In iterative settings, such as ML, the data owner will have to perform multiple mechanism executions, which necessitates the use of composition theorems to split the total guarantee into guarantees per iteration $(\varepsilon_i, \delta_i)$. Theorem 2 provides a bound for posterior belief β for DP under sequential composition, so, for k steps the data owner can simply divide $\varepsilon_i = \varepsilon/k$ and $\delta_i = \delta/k$, confirming that $\mathcal{A}_{\mathtt{DI}}$'s belief composes as expected by sequential composition. However, sequential composition only offers loose bounds in practice [DRV10, KOV17]. In addition, Theorem 3 states the bound ρ_α on membership advantage for a single mechanism execution without considering composition. We suggest to compose mechanisms with RDP composition, which permits a tight analysis of the privacy loss over a series of mechanisms. Therefore, we adapt both ρ_β and ρ_α to RDP.

(Maximum Posterior Belief) We first demonstrate that RDP composition results in stronger (ε, δ) guarantees than sequential composition for a fixed bound ρ_β:

$$\beta_k(\mathcal{D}|R) = \frac{1}{1 + \frac{\prod_{i=1}^k \Pr(\mathcal{M}_i(\mathcal{D}')=\vec{r}_i)}{\prod_{i=1}^k \Pr(\mathcal{M}_i(\mathcal{D})=\vec{r}_i)}} \tag{4.6}$$

$$\leq \frac{1}{1 + \prod_{i=1}^k e^{-(\varepsilon_{RDP,i} + (\alpha-1)^{-1}\ln(1/\delta_i))}} \tag{4.7}$$

$$= \frac{1}{1 + e^{k(\alpha-1)^{-1}\ln(1/\delta_i) - \sum_{i=1}^k \varepsilon_{RDP,i}}}$$

$$= \frac{1}{1 + e^{(\alpha-1)^{-1}\ln(1/\delta_i^k) - \sum_{i=1}^k \varepsilon_{RDP,i}}}$$

$$= \sum_{i=1}^k \varepsilon_{RDP,i} - (\alpha-1)^{-1}\ln(1/\delta_i^k) = \rho_\beta \tag{4.8}$$

Equation (4.8) implies that an RDP-composed bound can be achieved with a composed δ equal to δ^k. We know that sequential composition results in a composed δ value equal to $k\delta$.

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.

(Expected Membership Advantage) A similar analysis of the expected membership advantage under composition is required when considering a series of mechanism $\mathcal{M}$ and function f. We restrict our elucidations to the Gaussian mechanism. The k-fold composition of $\mathcal{M}_{Gau_i}$, each step guaranteeing $(\alpha, \varepsilon_{RDP,i})$-RDP, can be represented by a single execution of $\mathcal{M}_{Gau}$ with k-dimensional output guaranteeing $(\alpha, \varepsilon_{RDP} = k\varepsilon_{RDP,i})$-RDP. To prove this, we bound $\|\mu_{1,i} - \mu_{2,i}\|$ for each of the composed mechanism executions by Δf_2. Theorem 3 yields

$$
\begin{aligned}
\mathrm{Adv}^{\mathrm{DI}} &= 2\Phi(\Delta/2) - 1 \\
&= 2\Phi\left(\frac{\|\vec{\mu}_1 - \vec{\mu}_2\|_2}{2\sigma_i}\right) - 1 \\
&= 2\Phi\left(\frac{\sqrt{k}\|\mu_{1,i} - \mu_{2,i}\|_2}{2\Delta f_2\sqrt{\alpha/(2\varepsilon_{RDP,i})}}\right) - 1 \\
&\leq 2\Phi\left(\frac{\sqrt{k}}{2\sqrt{\alpha/(2\varepsilon_{RDP,i})}}\right) - 1 \\
&= 2\Phi\left(\sqrt{\frac{k\varepsilon_{RDP,i}}{2\alpha}}\right) - 1 \\
&= 2\Phi\left(\sqrt{\frac{\varepsilon_{RDP}}{2\alpha}}\right) - 1
\end{aligned}
$$

The result shows that $\mathcal{A}_{\mathrm{DI}}$ fully takes advantage of the RDP composition properties of $\varepsilon_{RDP,i}$ and α. As expected, ρ_a takes on the same value, regardless of whether k composition steps with $\varepsilon_{RDP,i}$ or a single composition step with ε_{RDP} is carried out. Therefore, we can calculate the final ρ_a for processes with multiple iterations, like the training of deep learning models, and a desired final ρ_a can be broken down into a privacy guarantee per composition step with RDP.

5. Application to deep learning using DPSGD

In DP, the use of global sensitivity Δf often results in a mechanism that yields unnecessarily high noise, which does not reflect the function's insensitivity to individual inputs [NRS07]. Because ε is often not a tight bound under global sensitivity, Nissim et al. [NRS07] proposed local sensitivity, which depends not only on the function, but also on the input data to be used. Local sensitivity no longer protects against inference on any possible adjacent datasets, but only on the chosen true dataset $\mathcal{D}$ and any dataset adjacent to it. The local sensitivity approach decreases noise addition by narrowing the guarantee, while still protecting the true dataset used for a calculation. We suggest a heuristic for estimating local sensitivity in a deep learning setting and investigate the impacts of this local sensitivity estimate on posterior belief and membership advantage. Here, a neural network (NN) is provided a training dataset $\mathcal{D}$ to learn a prediction function $\hat{y} = f_{nn}(\vec{x})$ given $(\vec{x}, y) \in \mathcal{D}$. Learning is achieved by means of an optimizer and a variety of differentially private optimizers for deep learning are available[1]. These optimizers represent a differentially private training algorithm that updates the weights θ_t per training step $t \in T$ with $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \tilde{g}_t$, where $\alpha > 0$ is the learning rate and $\tilde{g}_t = \mathcal{M}_t(\mathcal{D})$ denotes the Gaussian perturbed gradient. After T update steps, where each update step is itself an application of the Gaussian mechanism, the algorithm outputs a differentially private weight matrix θ_T which is then used in the prediction function $f_{nn}(\cdot)$. Considering the evaluation of $f_{nn}(\cdot)$ given $(\vec{x}, y) \in \mathcal{D}$ as post-processing of the trained weights θ_T, we find that prediction $\hat{y} = f_{nn}(\vec{x})$ is (ε, δ)-differentially private too.

We assume that $\mathcal{A}_{\mathtt{DI}}$ desires to correctly identify the dataset with the correct representation of a record d when having the choice between $\mathcal{D}$ and $\mathcal{D}'$ that differ in d. Furthermore, $\mathcal{A}_{\mathtt{DI}}$ is assumed to possess the initial weights θ_0, the perturbed gradients $\tilde{g}_t$ after every epoch, the values of privacy parameters (ε, δ), and sensitivity $\Delta f_2 = \mathtt{C}$ equal to the clipping norm. There are two variations of DP: bounded and unbounded. In bounded DP, it holds that $|\mathcal{D}| = |\mathcal{D}'|$, which was the standard in this work so far. Differentially private deep learning optimizers such as the one utilized in this work consider unbounded DP as the standard case [MAE+18] in which $|\mathcal{D}| - |\mathcal{D}'| = 1$, whereas we considered bounded DP until now. We continue to consider bounded DP, but we will extend our arguments and experiments to unbounded DP within this section. In both cases, $\mathcal{D}$ and $\mathcal{D}'$ are independently sampled from a distribution in line with Definition 1. In machine learning, the finite set *DOM* from which $\mathcal{D}$ and $\mathcal{D}'$ are sampled is the total available dataset. The datapoints not sampled will result in a test set, which can be used to evaluate the utility of the trained model.

In this section, Δf_2 refers to the sensitivity with which noise added by a mechanism is scaled, not necessarily global sensitivity. In some experiments, for example, we set $\Delta f_2 = LS_{\hat{g}_t}(\mathcal{D})$ with average clipped gradient calculation $\hat{g}_t$, using local sensitivity as in Definition 3, rather than global sensitivity as in Definition 2. The assumptions are very similar to those of white-box membership inference attacks for federated learning [NSH18]. In federated learning, multiple participants,

[1] All experiments within our work were realized by using the Tensorflow privacy package: `https://github.com/tensorflow/privacy`

each with their own private training data, jointly train a global model by sharing gradients for their subsets of training data with an aggregator, who combines the gradients and shares the aggregated update with all participants. In the role of either aggregator or participant, the adversary observes the updates of the model, as assumed in our experiments. The attack is defined in Algorithm 1. To implement bounded DP with global sensitivity, $\mathcal{D}'$ contains $n-1$ records and $\Delta f_2 = 2\mathtt{C}$, since the maximum influence of one example on the sum of per-example gradients is $\mathtt{C}$. If one record is replaced with another, the lengths of the clipped gradients of these two records could each be $\mathtt{C}$ and point in opposite directions, which results in $n \cdot \|\hat{g}_t(\mathcal{D}') - \hat{g}_t(\mathcal{D})\|_2 = 2\mathtt{C}$. We also note that the same value of Δf_2 used by $\mathcal{A}_{\mathtt{DI}}$ must also be used by $\mathcal{M}$ to add noise.

The motivation behind Algorithm 1 is intuitively as follows. The initial weights θ_0 and the clipping norm $\mathtt{C}$ can be thought of as constants in the gradient functions $\hat{g}_0(\mathcal{D})$ and $\hat{g}_0(\mathcal{D}')$, which also depend on cost function J. $\mathcal{A}_{\mathtt{DI}}$ computes these gradient values based on $J, \mathtt{C}, \theta_0, \mathcal{D}$, and $\mathcal{D}'$ and then compares them to the perturbed gradient $\tilde{g}_0$ to determine posterior belief β_1. $\mathcal{A}_{\mathtt{DI}}$ applies $\tilde{g}_0$ to θ_0 with knowledge of α to receive θ_1 and consequently repeats the cycle for all epochs, updating the posterior belief at every step.

Algorithm 1 Strong Adaptive Adversary in Deep Learning

Require: Datasets $\mathcal{D}$ and $\mathcal{D}'$ with n and $n-1$ records $\mathcal{D}_i$ and $\mathcal{D}'_i$, respectively, training steps T, cost function $J(\theta)$, perturbed gradients $\widetilde{g}_t$ for each training step $t \leq T$, initial weights θ_0, prior beliefs $\beta_0(\mathcal{D}) = \beta_0(\mathcal{D}') = 0.5$, learning rate α, clipping threshold C, and mechanism $\mathcal{M}$

Ensure: Adversary Confidence $\beta_T(\mathcal{D})$

for $t \in [T]$ **do Compute gradients**

For each $i \in \mathcal{D}, \mathcal{D}'$, compute $g_t(\mathcal{D}_i) \leftarrow \bigtriangledown_{\theta_t} J(\theta_t, \mathcal{D}_i)$ and $g_t(\mathcal{D}'_i) \leftarrow \bigtriangledown_{\theta_t} J(\theta_t, \mathcal{D}'_i)$

Clip gradients

Clip each $g_t(\mathcal{D}_i), g_t(\mathcal{D}'_i)$ for $i \in \mathcal{D}, \mathcal{D}'$ to have a maximum L^2 norm $\mathtt{C}$ using $\bar{g}_t(\mathcal{D}_i) \leftarrow g_t(\mathcal{D}_i)/max(1, \frac{\|g_t(\mathcal{D}_i)\|_2}{\mathtt{C}})$ and $\bar{g}_t(\mathcal{D}'_i) \leftarrow g_t(\mathcal{D}'_i)/max(1, \frac{\|g_t(\mathcal{D}'_i)\|_2}{\mathtt{C}})$

Calculate Batch gradients

$\hat{g}_t(\mathcal{D}) \leftarrow avg(\bar{g}_t(\mathcal{D}_i))$

$\hat{g}_t(\mathcal{D}') \leftarrow avg(\bar{g}_t(\mathcal{D}'_i))$

Calculate Sensitivity

$\Delta f_t \leftarrow \mathtt{C}$

Calculate Belief

$\beta_{t+1}(\mathcal{D}) \leftarrow \frac{\beta_t(\mathcal{D}) * Pr[\mathcal{M}(\hat{g}_t(\mathcal{D})) = \widetilde{g}_t]}{\beta_t(\mathcal{D}) * Pr[\mathcal{M}(\hat{g}_t(\mathcal{D})) = \widetilde{g}_t] + \beta_t(\mathcal{D}') * Pr[\mathcal{M}(\hat{g}_t(\mathcal{D}')) = \widetilde{g}_t]}$

Compute weights

$\theta_{t+1} \leftarrow \theta_t - \alpha \widetilde{g}_t$

end for

A motivating factor for MI attacks is that strong DP guarantees are difficult to achieve without significant utility loss in deep learning settings [JE19, BPS19]. This utility loss occurs because the DP guarantee protects all possible datasets, although only the training data itself must be protected in deep learning. Clipping norm $\mathtt{C}$ bounds the influence of a single training example on training by clipping each per-example gradient to the chosen value of $\mathtt{C}$. Although this value bounds the influence of a single example on the gradient, this bound is loose, since it does not necessarily reflect the difference between the training dataset and possible neighboring datasets. Thus, the gradients may be far from $\mathtt{C}$. In DP, when sensitivity is set to a value larger than necessary, the guarantee ε is not reached, so our metrics ρ_α and ρ_β will not be reached either. If $n \cdot \|\hat{g}_t(\mathcal{D}) - \hat{g}_t(\mathcal{D}')\| \ll \mathtt{C}$,

adversary confidence $\beta(\mathcal{D})$ would be very small in every case when $\Delta f_2 = \mathtt{C}$, which is the case in most implementations of differentially private neural networks. This scenario can be thought of as using a global sensitivity of C, rather than local sensitivity. We suggest addressing this by fixing the dataset $\mathcal{D}$ and considering only datasets $\mathcal{D}'$ adjacent to this fixed $\mathcal{D}$; however, approximating local sensitivity for neural network 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 avoid this dilemma, we propose a metric, *dataset sensitivity*, in Definition 6 with which we strive to consider the $\mathcal{D}$' with the largest difference to $\mathcal{D}$ within the ML dataset in an effort to approximate local sensitivity.

Definition 6 (Dataset Sensitivity)**.** *For a given ML dataset $\mathcal{U}$ and neighboring datasets $\mathcal{D}$, $\mathcal{D}' \subseteq \mathcal{U}$ the dataset sensitivity $DS(\mathcal{D})$ with respect to a distance function is*

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

The motivation behind Definition 6 is based on the assumption that local sensitivity can be approximated as $LS_{\hat{g}_t}(\mathcal{D}) = n \cdot ||\hat{g}_t(\mathcal{D}) - \hat{g}_t(\hat{\mathcal{D}}')||$ where $DS(\mathcal{D}) = ||\mathcal{D} - \hat{\mathcal{D}}'||$. The simplification from local sensitivity to dataset sensitivity allows us to bypass the complex gradient calculations. Instead of scaling noise to an arbitrarily chosen C, for which we show in Figure 5.1(b) that it is not necessarily tight, we calculate $DS(\mathcal{D})$ in order to identify $\hat{\mathcal{D}}'$. Based on our assumption, we can then scale noise to the approximated value of $LS_{\hat{g}_t}(\mathcal{D}) = n \cdot ||\hat{g}_t(\mathcal{D}) - \hat{g}_t(\hat{\mathcal{D}}')||$ and achieve local sensitivity for any weights and architecture. Our experimental observations further confirmed this expectation.

This procedure makes our choice of $\mathcal{D}$ indistinguishable from any $\mathcal{D}'$ within the chosen range of possible values. For our experiments, this range of values is the entire Modified National Institute of Standards and Technology database (MNIST) of handwritten digits, and $\mathcal{D}$ and $\mathcal{D}'$ each contain 100 data points. We choose the data points x_1 and x_2 in MNIST that result in the highest dataset sensitivity DS($\mathcal{D}$), with $x_1 \in \mathcal{D}$ and $x_2 \notin \mathcal{D}$. $\mathcal{D}'$ is then formed by replacing x_1 with x_2 in $\mathcal{D}$. We measure similarity for the dataset sensitivity with the structural similarity index measure (SSIM)[2], which allows us to examine individual pairs of datapoints instead of the entire datasets because $||\mathcal{D} - \mathcal{D}'|| = ||x_1 - x_2||$. Since different images will result in very different gradients, we can calculate $DS(\mathcal{D}) = ||x_1 - x_2||$ and approximate $LS_{\hat{g}_t}(\mathcal{D}) = n \cdot ||\hat{g}_t(\mathcal{D}) - \hat{g}_t(\mathcal{D}')|| = ||\hat{g}_t(x_1) - \hat{g}_t(x_2)||$. For unbounded DP, we remove the data point with the smallest SSIM distance to all other images from $\mathcal{D}$ to form $\mathcal{D}'$.

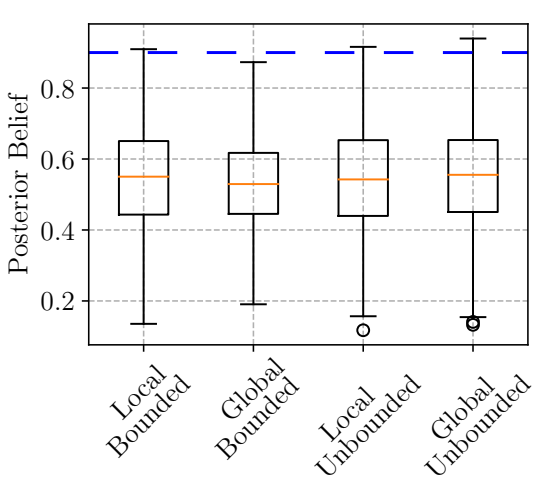
Based on the previously introduced assumptions and notations we adapt $\mathcal{A}_{\mathtt{DI}}$ to local sensitivity. The implementation of $\mathcal{A}_{\mathtt{DI}}$ for the differentially private stochastic gradient descent is stated in Algorithm 1 and specifies $\mathcal{A}_{\mathtt{DI}}$ in an unbounded environment with global sensitivity. For bounded DP with local sensitivity, Algorithm 1 can be adjusted, s.t. $\mathcal{D}'$ is fixed to n training records, and $\Delta f_2 = LS_{\hat{g}_t}(\mathcal{D}) = n \cdot \|\hat{g}_t(\mathcal{D}') - \hat{g}_t(\mathcal{D})\|_2$. To implement unbounded DP with local sensitivity, $\Delta f_2 = n \cdot \|n \cdot \hat{g}_t(\mathcal{D}') - (n-1) \cdot \hat{g}_t(\mathcal{D})\|_2$ and $\mathcal{D}'$ contains $n-1$ records.

5.1 Setting of the experiment

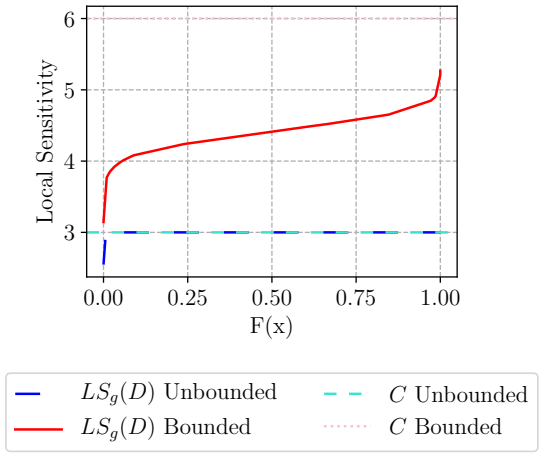
For practical evaluation, we build a feed-forward NN for the MNIST dataset[3]. For MNIST our NN architecture consists of two repetitions of a convolutional layer with kernel size $(3,3)$, batch

[2] `https://ece.uwaterloo.ca/~z70wang/research/ssim/`

[3] Overview and detailed description available at: `http://yann.lecun.com/exdb/mnist/`



(a) Confidence distribution



(b) Measured $LS_{\hat{g}_t}(\mathcal{D})$ in bounded and unbounded settings

Figure 5.1: Sensitivity and posterior belief (30 epochs) for $\rho_\beta = 0.9$ and $\delta = 0.01$

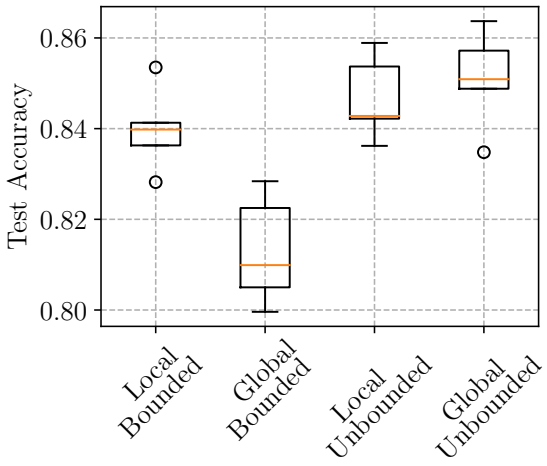


Figure 5.2: Distribution of test accuracy (30 epochs) for $\rho_\beta = 0.9$ and $\delta = 0.01$

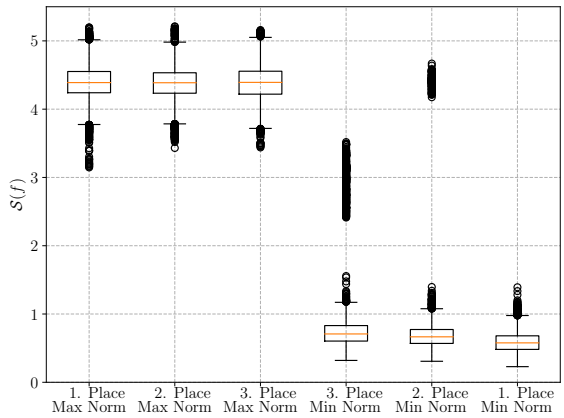


Figure 5.3: Distribution of $n \cdot ||\hat{g}_t(\mathcal{D}) - \hat{g}_t(\mathcal{D}')||$ from max to min difference in $\mathcal{D}$ and $\mathcal{D}'$

	Local Δf_2	Global Δf_2
Bounded DP	$(0.239, 0.002)$	$(0.194, 0)$
Unbounded DP	$(0.228, 0.002)$	$(0.273, 0.004)$

Table 5.1: Empirical advantage (α, δ) and empirical δ which is within δ

normalization and max pooling with pool size $(2,2)$ before being flattened for the output layer. We use *relu* and *softmax* activation functions for the convolutional layers and the output layer, respectively.

One epoch represents the evaluation of all records in $\mathcal{D}$. Thus, it is important to highlight that the number of update steps T varies in practice depending on the number of records from $\mathcal{D}$ used for calculating the DP gradient update $\tilde{g}$. In *mini-batch gradient descent* a number of b records from $\mathcal{D}$ is used for calculating an update and one epoch results in $t = \mathcal{D}/b$ update steps. In contrast in *batch gradient descent* all records in $\mathcal{D}$ are used for calculating the update and each epoch consists of a single update step. While all approaches vary in their speed of convergence due to the gradient update behavior (i.e., many small updates vs. few large updates) none of the approaches has hard limitations w.r.t. convergence of accuracy and loss. Within our work, we operate with batch gradient descent and given differentially private gradient updates $\tilde{g}$ after any update step t $\mathcal{A}_{\mathtt{DI}}$ decides whether $\tilde{g}$ was calculated on $\mathcal{D}$ or $\mathcal{D}'$. We assume that $\mathcal{A}_{\mathtt{DI}}$ has equal prior beliefs of 0.5 on $\mathcal{D}$ and $\mathcal{D}'$. The prior belief of $\mathcal{A}_{\mathtt{DI}}$ adapts at every step t.

In the experiments, relevant parameters are set as follows: training data $|\mathcal{D}| = 100$, epochs $k = 30$, clipping norm $\mathtt{C} = 3.0$, learning rate $\alpha = 0.005$, $\delta = 0.01$, and $\rho_\beta = 0.9$. The overall privacy parameter ε, transformed from ρ_β, is split into a per step privacy parameter according to RDP composition. We choose $\mathcal{D}$ randomly and select $\mathcal{D}'$ to maximize dataset sensitivity. This choice of $\mathcal{D}'$ represents the case in which the bounds will be most closely reached, since the sensitivity between $f(\mathcal{D}) and f(\mathcal{D}')$ is actually reached. In practice, during private deep learning, $\tilde{g}$ must be calculated also for $\mathcal{D}'$ if local sensitivity is to be estimated; otherwise, $\tilde{g}$ on $\mathcal{D}'$ does not have to be computed. We merely do so in order to empirically validate $\mathcal{A}_{\mathtt{DI}}$ during training. We run experiments for both local and global sensitivity to evaluate the effect of the sensitivity on identifiability and utility, and we evaluate both bounded and unbounded settings. For each of the four cases, we train a neural network and simulate $\mathcal{A}_{\mathtt{DI}}$ 1000 times. We then analyze the resulting distribution of posterior beliefs and calculate the membership advantage by counting the cases in which posterior belief for $\mathcal{D}$ exceeds 0.5.

5.2 Results

We present the empirically calculated values (i.e., after the training) for $\rho_a = 0.2562$ and δ in Table 5.1. The belief distributions for the described experiments can be found in Figure 5.1(a).

Note that although Figure 5.1(a) shows posterior belief $\beta_T(\mathcal{D})$ exceeding ρ_β in some cases, Table 5.1 confirms that δ indeed bounds the percentage of experiments for which $\beta_T(\mathcal{D}) > \rho_\beta$. For all experiments with local sensitivity and for global, unbounded DP, the empirical and analytical values of ρ_a match the empirical values. However, in global, bounded differential privacy the difference of correct guesses and incorrect guesses by $\mathcal{A}_{\mathtt{DI}}$ falls below ρ_a. The percentage of evaluation runs for which $\beta_T(\mathcal{D}) > \rho_\beta$ is also far lower. This behavior confirms the hypothesis that $\mathtt{C}$ is loose, so global sensitivity results in a lower value of $\beta_T(\mathcal{D})$, as is again confirmed by Figures 5.1(a) and 5.1(b). We also notice that the distributions in Figures 5.1(a) for local sensitivity

in bounded and unbounded settings look identical to each other. This observation confirms that the strong adaptive adversary attack model is applicable to chose privacy parameter ε in deep learning.

We now investigate the reason for the similarities between unbounded differential privacy with local and global sensitivity and also for the differences between Figures 5.1(a) concerning bounded differential privacy with local and global sensitivity. In the unbounded case, the distributions seem identical, which occurs when $\Delta f_2 = LS_{\hat{g}_t}(\mathcal{D}) = \|(n-1)\cdot\hat{g}_t(\mathcal{D}') - n\cdot\hat{g}_t(\mathcal{D})\|_2 = \mathtt{C}$, so the clipped per example gradient of the differentiating example in $\mathcal{D}$ should have the length 3, which is equal to $\mathtt{C}$. This hypothesis is confirmed with a glance at the development of $\|(n-1)\cdot\hat{g}_t(\mathcal{D}') - n\cdot\hat{g}_t(\mathcal{D})\|_2$ in Figure 5.1(b). This behavior is not surprising, since all per example gradients over the course of all epochs were greater than or close to $\mathtt{C} = 3$. In the bounded differential privacy experiments, $\Delta f_2 = LS_{\hat{g}_t}(\mathcal{D}) = n\cdot\|\hat{g}_t(\mathcal{D}') - \hat{g}_t(\mathcal{D})\|_2 \neq 2\mathtt{C}$, since the corresponding distributions in Figure 5.1(a) do not look identical. This expectation is confirmed by the plot of $n\cdot\|\hat{g}_t(\mathcal{D}') - \hat{g}_t(\mathcal{D})\|_2$ in Figure 5.1(b). This difference implies that the per example gradients of the differentiating examples in $\mathcal{D}'$ and $\mathcal{D}$ are less than $2\mathtt{C}$ and do not point in opposite directions. We also point out that the length of gradients tends to decrease over the course of training, so if training converges to a point in which gradients are shorter than the chosen value of $\mathtt{C}$, globally differentially private deep learning inherently offers a stronger privacy guarantee than was originally chosen.

A glance at Figure 5.2 confirms that the differentially trained models in these models do, indeed, yield some utility. The visualized accuracy was achieved by increasing the training set size to $10,000$. We also observe that test accuracy is directly affected by the value of sensitivity Δf_2 chosen for noise addition. Since gradients in all four scenarios are clipped to the same value of $< \mathtt{C}$, the only differences between training the neural networks is Δf_2. As visualized in Figure 5.1(b), global and local sensitivities for unbounded DP were identical, so the nearly identical corresponding distributions in Figure 5.2 do not come as a surprise. Similarly, we observe that Δf_2 is greater for global, bounded DP in Figure 5.1(b), so utility is also lower for this case in Figure 5.2. The unbounded DP case with local sensitivity yields the highest utility, which can be explained by the low value of Δf_2 that can be read from Figure 5.1(b).

To confirm our claim that maximizing dataset sensitivity from Definition 6 allows us to approximate local sensitivity, we carry out network training for 250 runs with differing choices of $\mathcal{D}'$. We first evaluated the top three choices of $\mathcal{D}'$ that maximize dataset sensitivity, then the three choices that minimize dataset sensitivity most. The resulting $n\cdot||\hat{g}_t(\mathcal{D}) - \hat{g}_t(\mathcal{D}')||$ can be read from Figure 5.3. We see that the choices which maximize dataset sensitivity result in larger values, while choosing $\mathcal{D}'$ to minimize dataset sensitivity results in a smaller value. We therefore see a downward trend from left to right in Figure 5.3.

6. Conclusions

We defined two identifiability bounds for the strong DP adversary in Machine Learning with the differentially private stochastic gradient descent, maximum posterior belief ρ_β and expected membership advantage ρ_a. These two bounds can be transformed to privacy parameters (ε, δ). In consequence, with ρ_a and ρ_β, data owners and data scientists can map legal and social expectations towards identifiability to corresponding privacy parameters (ε, δ). Furthermore, we implemented an instance of the DP adversary for Machine Learning and showed that the bounds can be reached under multidimensional queries with composition. To reach the bound it is necessary that the sensitivity is reflecting the actual local sensitivity of the dataset. We approximate $LS_f(\mathcal{D})$ for stochastic gradient descent, improving the utility of the differentially private model training when compared to the use of global sensitivity Δf and reaching the bounds.

Within MOSAICrOWN, and in data markets after the project concluded, we see large potential for choosing the privacy parameter as a transformation of identifiability bounds due to two reasons. First, identifiability bounds are on a well-defined scale between 0.5 and 1.0 and therefore go along with probabilities that we encounter and assess already in our daily life. In contrast privacy parameter ε is defined between 0 and positive infinity. Second, if paired with local sensitivity our bounds can actually be reached and thus, higher privacy parameters ε might be usable. This will likely have a positive effect on utility, which we illustrated in our evaluation.

Bibliography

[ACG+16] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep Learning with Differential Privacy. In *Proceedings of the Conference on Computer and Communications Security*, (CCS), New York, NY, USA, 2016. ACM Press.

[Ame10] American Department of Health and Human Services. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (HIPAA) privacy rule. `https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html`, 2010.

[AS19] John M. Abowd and Ian M. Schmutte. An economic analysis of privacy protection and statistical accuracy as social choices. *American Economic Review*, 109(1), 2019.

[BGRK19] Daniel Bernau, Philip-William Grassal, Jonas Robl, and Florian Kerschbaum. Assessing differentially private deep learning with membership inference. *arXiv preprint arXiv:1912.11328*, 2019.

[BPS19] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Proceedings of the Conference on Advances in Neural Information Processing Systems*, NIPS, Red Hook, NY, USA, 2019. Curran Associates Inc.

[CK12] Andrew Chin and Anne Klinefelter. Bounds on the sample complexity for private learning and private data release. *North Carolina Law Review*, 90(5), 2012.

[CT13] Chris Clifton and Tamir Tassa. On syntactic anonymity and differential privacy. In *Proceedings of the International Conference on Data Engineering Workshops*, ICDEW, New York, NY, USA, 2013. IEEE Computer Society.

[CYZF20] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A Taxonomy of Membership Inference Attacks against Generative Models. In *Proceedings of the Conference on Computer and Communications Security*, CCS, 2020.

[DKM+06] Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In *Proceedings of the International Conference on the Theory and Applications of Cryptographic Techniques*, EUROCRYPT, Berlin, Heidelberg, 2006. Springer.

[DR14] Cynthia Dwork and Aaron Roth. The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), 2014.

[DR16] Cynthia Dwork and Guy N. Rothblum. Concentrated Differential Privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[DRV10] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and Differential Privacy. In *Proceedings of the Symposium on Foundations of Computer Science*, (FOCS), Piscataway, NJ, USA, 2010. IEEE Computer Society.

[Dwo06] Cynthia Dwork. Differential Privacy. In *Proceedings of the International Colloquium on Automata, Languages and Programming*, ICALP, Berlin, Heidelberg, 2006. Springer.

[Eur16] European Parliament and Council of the European Union. General data protection regulation. *Official Journal of the European Union*, 119(1), April 2016.

[HGH+14] Justin Hsu, Marco Gaboardi, Andreas Haeberlen, Sanjeev Khanna, Arjun Narayan, Benjamin Pierce, and Aaron Roth. Differential privacy: An economic method for choosing epsilon. In *Proceedings of the Computer Security Foundations Workshop*, CSFW, Piscataway, NJ, USA, 2014. IEEE Computer Society.

[HMDD19] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies (PoPETs)*, 2019(1), 2019.

[JE19] Bargav Jayaraman and David Evans. Evaluating differentially private machine learning in practice. In *Proceedings of the USENIX Security Symposium*, (SEC), Berkeley, CA, USA, 2019. USENIX Association.

[JWK+20] Bargav Jayaraman, Lingxiao Wang, Katherine Knipmeyer, Quanquan Gu, and David Evans. Revisiting Membership Inference Under Realistic Assumptions. *arXiv preprint arXiv:2005.10881*, 2020.

[KM11] Daniel Kifer and Ashwin Machanavajjhala. No free lunch in data privacy. In *Proceedings of the International Conference on Management of Data*, SIGMOD, New York, USA, 2011. ACM Press.

[KOV17] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The Composition Theorem for Differential Privacy. *IEEE Transactions on Information Theory*, 63(6), 2017.

[LC11] Jaewoo Lee and Chris Clifton. How much is enough? choosing epsilon for differential privacy. In *Proceedings of the International Conference on Information Security*, ISC, Berlin, Heidelberg, 2011. Springer.

[LC12] Jaewoo Lee and Chris Clifton. Differential identifiability. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, KDD, New York, NY, USA, 2012. ACM Press.

[LQS+13] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the Conference on Computer and Communications Security*, CCS, New York, NY, USA, 2013. ACM Press.

[MAE+18] H. Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. A general approach to adding differential privacy to iterative training procedures. *arXiv preprint arXiv:1812.06210*, 2018.

[Mah36] Prasanta C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, 2(1), 1936.

[Miv17] Ilya Mivonov. Rényi Differential Privacy. In *Proceedings of the IEEE Computer Security Foundations Symposium*, CSF, Piscataway, NJ, USA, 2017. IEEE Computer Society.

[MKB79] Kanti V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate analysis*. Academic Press London, New York, NY, USA, 1979.

[Nis16] Helen Nissenbaum. Differential Privacy in Context: Conceptual and Ethical Considerations. `https://www.ias.edu/ideas/2016/differential-privacy-symposium`, 2016.

[NRS07] Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the Symposium on Theory of Computing*, STOC, New York, USA, 2007. ACM Press.

[NSH18] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Stand-alone and federated learning under passive and active white-box inference attacks. *arXiv preprint arXiv:1812.00910*, 2018.

[NW18] Kobbi Nissim and Alexandra Wood. Is privacy privacy? *Philosophical Transactions of the Royal Society*, 376(2128), 2018.

[OK20] Daniel Oberski and Frauke Kreuter. Differential Privacy and Social Science: An Urgent Puzzle. *Harvard Data Science Review*, 2(1), 2020.

[Par14] Article 29 Data Protection Working Party. Opinion 05/2014 on anonymisation techniques. `https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf`, 2014.

[RRLM18] Md. Atiqur Rahman, Tanzila Rahman, Robert Laganière, and Noman Mohammed. Membership inference attack against differentially private deep learning model. *Transactions on Data Privacy*, 11, 2018.

[Sam01] Pierangela Samarati. Protecting respondents' identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 13(6), November 2001.

[SS98] Pierangela Samarati and Latanya Sweeney. Generalizing data to provide anonymity when disclosing information. In *Proceedings of the Symposium on Principles of Database Systems*, PODS, 1998.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the Symposium on Security and Privacy*, S&P, Piscataway, NJ, USA, 2017. IEEE Computer Society.

[TG96] Kagan Tumer and Joydeep Ghosh. Estimating the bayes error rate through classifier combining. In *Proceedings of the International Conference on Pattern Recognition*, volume 2 of *ICPR*, Piscataway, NJ, USA, 1996. IEEE Computer Society.

[vEH10] Tim van Erven and Peter Harremoës. Rényi divergence and majorization. In *Proceedings of the Symposium on Information Theory*, ISIT, Piscataway, NJ, USA, 2010. IEEE Computer Society.

[YGFJ18] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. *arXiv preprint arXiv:1709.01604*, 2018.