# Analysis and Measurement of Attack Resilience of Differential Privacy

Patricia Guerra-Balboa
patricia.balboa@kit.edu
Karlsruhe Institute of Technology,
KASTEL Security Research Labs
Karlsruhe, Germany

Annika Sauer
annika.sauer@student.kit.edu
Karlsruhe Institute of Technology,
KASTEL Security Research Labs
Karlsruhe, Germany

Thorsten Strufe
thorsten.strufe@kit.edu
Karlsruhe Institute of Technology,
KASTEL Security Research Labs
Karlsruhe, Germany

## Abstract

Differential Privacy (DP) is the de facto standard privacy metric in private learning. Its robust mathematical definition makes it especially appealing for global data analytics without compromising individual privacy.

However, DP resilience against state-of-the-art attacks is not formalized consistently, and the interpretation of the privacy implications of parameter choices is not intuitive. This formalization is relevant because DP relies on the choice of a privacy budget, which is crucial in obtaining a good trade-off between the privacy of the individuals in the dataset and the utility of the results for data analysis.

This paper presents a systematic overview of theoretical bounds obtained in the literature on DP resilience against three types of attacks: the membership-inference, the attribute-inference, and the data reconstruction attacks. For each attack, we introduce tighter theoretical bounds and analyze the limitations of existing performance metrics. To overcome these limitations, we propose a new attack performance metric: Unbiased Reconstruction Robustness. In addition, we prove the relation between Unbiased Reconstruction Robustness and the already existing metrics, showing its consistency. Finally, we prove a new bound for this metric in the presence of DP.

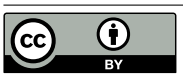
## CCS Concepts

• **Mathematics of computing** → *Probabilistic algorithms*; *Hypothesis testing and confidence interval computation*; • **Security and privacy** → **Data anonymization and sanitization**.

## Keywords

Differential Privacy, Attack Resilience, Adversarial Bound, Private Machine Learning

**ACM Reference Format:**
Patricia Guerra-Balboa, Annika Sauer, and Thorsten Strufe. 2024. Analysis and Measurement of Attack Resilience of Differential Privacy. In *Proceedings of the 23rd Workshop on Privacy in the Electronic Society (WPES '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3689943.3695046





*WPES '24, October 14–18, 2024, Salt Lake City, UT, USA*

ACM ISBN 979-8-4007-1239-5/24/10
https://doi.org/10.1145/3689943.3695046

## 1 Introduction

While data analytics applications improve economic and societal welfare, they also pose an increasing risk to individual privacy. Several real-world attacks on private data have shown the risk of publishing pseudonymized data [30, 26, 6, 7, 9, 31]. To deal with the conflicting goals of extracting data statistics without harming individuals' private information, several techniques and privacy notions have been proposed [19]. Among them, Differential Privacy (DP) has been established as the standard notion for privacy-preserving data analysis [8], particularly in the field of private learning [13, 28, 20].

DP allows us to learn statistics about the population while providing strict privacy guarantees. The privacy guarantee of DP is parameterized by the *privacy budget* $\varepsilon$. The privacy budget controls how similar the probabilities of observing the same output are, independent of whether an individual participated in the data collection or not [13]. The choice of this parameter is key. If $\varepsilon$ is too large, private information will be disclosed. Choosing $\varepsilon$ too small can significantly reduce the usefulness of the mechanism's output for data analysis [2].

DP has several desirable properties, such as the assumption of a strong adversary, robustness to post-processing, and composability [13]. However, its weak point compared to other notions is the lack of a precise connection between the privacy parameters and the interpretation of which privacy guarantees the user obtains [5, 23]. Among practitioners, there is no clear consensus on how to choose the privacy budget [11, 23], and any scientific work that seeks to explain the impact of the choice of $\varepsilon$ focuses on specific algorithms or settings (e.g., [25]).

The missing interpretation of DP parameters is particularly tangible in the study of attack resilience of DP learning mechanisms. We find three particular privacy attacks in the context of Machine Learning (ML): Membership Inference Attacks (MIAs) [29], Attribute Inference Attacks (AIAs) [35], and Data Reconstruction Attacks (DRAs) [1]. While these attacks can be applied to any inference process [36], they have been extensively studied in ML applications. Several empirical studies [2, 37, 34] show their success over unprotected ML algorithms and demonstrate a tendency of DP to be an effective mitigation strategy in practice. However, achieving good model performance usually requires choosing larger privacy budgets than is preferable from a formal perspective, and the exact relationship between chosen privacy budgets and achieved protection remains an open question. This complicates the application of DP as it is unclear what budget is required to prevent a particular attack.

Due to the need to understand the ability of attacks to infer private information about participants, various attack performance metrics have been introduced. An attack performance metric is a measurement tool used to quantify the effectiveness and impact of privacy attacks on a machine learning model. The particular case of MIAs has been widely studied, and the standard performance metric is the *membership advantage* [35]. Several bounds on the advantage of a MIA under DP exist [18, 35, 14]. However, other attacks, such as AIAs, have no general bounds on their performance under DP protection. In the case of DRAs, we find one attack performance metric, ReRo [1]. Balle et al. [1] provide a formal relationship between this metric and DP, establishing the only general criterion so far for quantifying the mitigation of arbitrary attack success due to DP, as we discuss in Section 4.

However, the relationship between $\varepsilon$ and ReRo proved in [1] is generally not tight, leading to an overestimation of privacy risk and thus an unnecessary loss of utility. Another problem is that ReRo does not have the same interpretation as the state-of-the-art performance metrics for MIAs and AIAs. Performance for these attacks is measured with advantages that quantify the amount of information (either about the membership or the attributes) of a target that is leaked specifically by participating in the training of the mechanism $\mathcal{M}$. ReRo, on the other hand, is a success probability. Therefore, the results of ReRo are difficult to compare with previous work since it measures a different aspect of the corresponding attack performance (see Section 4.3). In particular, we show in Section 6 that ReRo cannot distinguish whether an attack succeeds in reconstructing a participant's private information because this target information has been learned by the mechanism (privacy leakage) or because of some external source of information unrelated to participation in the training, such as the global distribution of the population (privacy fallacy). Hence, ReRo may overestimate the privacy risk and is not directly comparable to existing metrics. Thus, its ability to conclude which type of attack is particularly effective under which learning algorithms and parameter choices for DP is limited. Given that the choice of parameters directly affects the privacy-utility trade-off, this issue needs further investigation [2].

In this work, we establish a general framework for assessing the protection that DP provides against attacks on differentially private learning. This framework allows us to establish consistent comparisons between the protection against different attacks and a precise interpretation that can be used as a criterion for choosing the privacy parameters.

To this end, we perform a systematization of the adversarial bounds of DP provided in the state-of-the-art research for MIAs, AIAs, and DRAs. We provide interpretations and relationships for the existing attack performance metrics and the bounds on these performances derived from DP. We describe the current limitations, including the overestimation of privacy leakage and inconsistencies in the state-of-the-art research. In addition, we prove a tighter bound for $(0, \gamma)$-ReRo under DP protection, improving on the problem of the previous bound [1] that underestimates DP protection.

Finally, to overcome the discussed limitations, we propose a new attack performance metric: Unbiased Reconstruction Robustness (U-ReRo). We prove that DP bounds U-ReRo for arbitrary attacks and give a bound that relates U-ReRo directly to $\varepsilon$. Consequently, we provide U-ReRo as a useful criterion for $\varepsilon$-selection in DP learning that overcomes the problems of ReRo.

Compared to ReRo, U-ReRo does not lose any generality since it applies to any data distribution and attack against private learning. Additionally, it overcomes the problem of the privacy fallacy, as explained in Section 7. U-ReRo compares the probability that the adversary correctly reconstructs the target record when it is a record from the dataset with the probability of correctly reconstructing a record drawn from the underlying distribution. In this way, U-ReRo only gives the adversary credit for correctly reconstructing the participants when it fails to reconstruct non-participants. ReRo, on the other hand, can be arbitrarily large for an attack where the adversary's reconstruction ability is the same for participants and non-participants because the success comes from information unrelated to participation in training, such as the global distribution of the population. Therefore, U-ReRo outperforms the existing metric for estimating the actual privacy risk.

Since U-ReRo's estimate of privacy risk is more accurate than ReRo's, it allows for the selection of a higher $\varepsilon$. We show in Section 7 that U-ReRo provides a significantly lower risk value. Hence, by using our metric, practitioners can select larger $\varepsilon$ values, which imply less noise in DP mechanisms and thus more utility, while maintaining the same privacy risk.

Moreover, U-ReRo is not a success probability like ReRo but an advantage like the standard performance metrics for MIA (the membership advantage) and AIA (the attribute advantage). We show the consistency of U-ReRo with the previous advantages by proving that when applied to the particular scenario in which these metrics were defined, it yields the same values. We conclude that U-ReRo is a generalization of the widely used membership and attribute advantages, thus overcoming the inconsistency of ReRo.

The contributions of this paper are:

- We systematize the existing formal knowledge about DP attack resilience and expose the limitations of ReRo as an existing performance metric. We show the divergence between the interpretation of this metric and the other metrics in the literature, pointing out its overestimation of privacy leakage due to the privacy fallacy.
- We prove a tighter bound for $(0, \gamma)$-ReRo, allowing for better estimation and smaller consumption of the privacy budget.
- We define the new attack performance metric U-ReRo and prove a result that directly relates U-ReRo and $\varepsilon$. Our theorem provides a novel criterion for choosing $\varepsilon$ in DP learning. Compared to ReRo, it gives a more accurate assessment of the privacy risk of an individual's participation, leading to an improvement in the privacy-utility trade-off. Furthermore, we prove the formal relationship between U-ReRo and the attribute and membership advantages, thus showing its consistency.
- We use U-ReRo to prove a novel bound for the advantage of an arbitrary AIA under DP. We are the first in the literature to do so.

We provide the theorems and proof intuitions throughout the main body of this work. All formal proofs of our results can be found in Appendix A.

## 2 Related Work

Studies on attacks on private data under DP have a strong focus on ML. In this case, the adversary has access to a model trained on a sensitive database and possibly additional information (such as the distribution from which its training data was drawn). The attack objective is to infer private information about individuals in the training data.

The most studied privacy attacks against training data in the context of DP are MIAs. Their goal is to determine if a specific record is a member of the data used to train a given model. MIAs have been analyzed both empirically [2, 37, 34] and formally. We can find several theoretical bounds in literature (e.g., [35, 27, 18]).

AIAs are comparatively less explored and are usually studied from an empirical rather than a theoretical standpoint (e.g., [15, 33]). Their goal is to infer an attribute of an individual record that is assumed to be a member of the training data of a given model. Explicit theoretical bounds for this attack remain to be established. The only theoretical bounds for AIAs correspond to the particular scenario in which the AIA is either performed as a MIA [35] or is analyzed under special circumstances such as data dependencies [32]. In this work, we overcome this issue by proving a bound on the attribute advantage independently of the attack strategy.

Recently, Balle et al. [1] formulated DRAs as a generalization of possible attacks where the information inferred is compared to the ground truth with an error function. They introduce the concept of reconstruction robustness to analyze privacy leakage under reconstruction attacks and subsequently demonstrate that reconstruction robustness implies DP and vice versa. The results yield theoretical bounds for DP under DRAs. However, the bound provided for this metric is not generally tight. We improve on this by proving a tighter bound under perfect reconstruction. Furthermore, we show the intrinsic limitations of ReRo in terms of interpretability due to the privacy fallacy and its inconsistency with the membership and attribute advantages.

Other aspects of attacks on private data have been presented in the literature. One field of research focuses on the guarantees that DP can provide under data dependencies and demonstrates that established guarantees can be undermined if the entries in the dataset are correlated (e.g. [32, 37]). Another related research direction seeks implementation errors in DP algorithms that result in violations of the theoretical guarantees (e.g. [3, 10]). Finally, even if DP algorithms are implemented correctly, adversaries can still exploit side-effects like computation time and memory usage to gather information on a dataset [17]. These references present relevant threats to the attack resilience of DP. However, since these threats result from side effects like algorithmic implementation and data dependencies, they are beyond the scope of this paper.

## 3 Background

In this section, we formalize our understanding of DP in the context of private learning. A full overview of the notation used in this paper can be found in Table 1.

| Notation | Meaning |
|---|---|
| $z = (x, y) \in \mathcal{X} \times \mathcal{Y}$ | Data point with attributes $x$ and response $y$ |
| $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ | Data domain |
| $Z \sim \pi$ | Random variable following distribution of data records $\pi$ |
| $\pi^n$ | Distribution over datasets of size $n$ |
| $\mathcal{Z}^n$ | Space of datasets of size $n$ |
| $D$ | Dataset |
| $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ | Randomized learning mechanism |
| $\theta \sim \mathcal{M}(D)$ | Random variable following the distribution of output of the model trained on $D$ |
| $O \equiv \mathcal{M}_D$ | Output model trained on $D$ |
| $D_-$ | Training database missing one record |
| $\mathcal{M}_z \equiv \mathcal{M}(D_z)$ | Model trained on $D_- \cup \{z\}$ |
| $A$ | Attack |

**Table 1: Table of Notation.**

### 3.1 Differential Privacy

DP aims to hide the record of any participant in a database when an analyst extracts statistics about the whole population. It implies that the ability of an adversary to learn private information about individuals is limited. This limit on the inference abilities is established as follows:

*Definition 3.1 ($(\varepsilon, \delta)$-Differential Privacy [12]).* A randomized mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is $(\varepsilon, \delta)$-differentially private if for all $S \subseteq \Theta$ and for every pair of datasets $D, D' \in \mathcal{Z}^n$ such that $d_H(D, D') \leq 1$:

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

where $d_H(D, D')$ denotes the Hamming distance.

The Hamming distance [24] returns the number of records that need to be changed to transform $D$ into $D'$. Two databases $D$ and $D'$ are called *bounded-neighboring*, if it holds that $d_H(D, D') = 1$. Note that variations of DP under other neighborhood definitions exist. For instance, if we choose the symmetric difference between multisets instead of the Hamming distance, we obtain *unbounded DP* [22]. In this paper, we will only consider bounded DP.

Informally, given a DP mechanism $\mathcal{M}$, Definition 3.1 tells us that the probability that $\mathcal{M}(D)$ outputs $O$ is very close to the probability that $\mathcal{M}(D')$ outputs $O$ for any pair of neighboring databases, where "very close" is determined by $\varepsilon$. Thus, it cannot be easily distinguished whether $\mathcal{M}$ has been executed on $D$ or $D'$ based on its output. Since $D$ and $D'$ only differ in one record, it follows that the statistical significance of any inference about that record is bound.

The bound on the inference ability is not a binary attribute but a parameterized one. The parameters are the privacy budget $\varepsilon$ and the parameter $\delta$. The privacy budget controls the level of indistinguishability between outputs provided by mechanism $\mathcal{M}$. A lower $\varepsilon$ characterizes a stronger guarantee since the probabilities of observing $O$ under $D$ and observing $O$ under $D'$ are closer.

The parameter $\delta$ introduces the possibility that the privacy guarantee postulated by the privacy budget $\varepsilon$ may not hold in all cases. Consequently, $\delta$ allows certain violations of $\varepsilon$-DP while characterizing how likely such failures are to occur. If we do not allow any violations and $\delta = 0$, we speak of *pure DP*.

### 3.2 Learning algorithms in the context of DP

We introduce here the basic notation and concepts from learning theory and its relation with DP following [1].

Let $z = (x, y) \in \mathcal{X} \times \mathcal{Y} \equiv \mathcal{Z}$ be a data record where $x$ represents a set of features or attributes and $y$ a response. We consider a learning algorithm $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ such that, given a training database $D \in \mathcal{Z}^n$, it produces a trained model $O = \mathcal{M}_D \in \Theta$. By abuse of notation, we will denote $\mathcal{M}(D)$ as the distribution of output models when trained in $D$, and $\theta \sim \mathcal{M}(D)$ as the random variable.

The goal of $\mathcal{M}$ is to produce a trained model $\mathcal{M}_D$ that approximately minimizes the expected value of a loss function $l$ over $D$. If $z = (x, y) \in D$, the loss function $l(\mathcal{M}_D(x), z)$ measures how much $\mathcal{M}_D(x)$ differs from $y$.

We assume an underlying distribution of data records $\pi$, such that each record $z$ is drawn independently from $Z \sim \pi$. Therefore, the training data follows the distribution $\pi^n$.

We consider the training data $D$ as the target of the attack. A training algorithm $\mathcal{M}$ is $\varepsilon$-DP when for all training databases $D, D'$ such that $d_H(D, D') \leq 1$ and for all possible training algorithms $O = \Theta$ it holds that:

$$\Pr(\mathcal{M}(D) = O) \leq e^{\varepsilon} \Pr(\mathcal{M}(D') = O).$$

The protected output is the combination of all training steps, in other words, the trained model. This model can subsequently be queried or analyzed in other ways at liberty. Here we make an abuse of notation that will hold for the rest of the paper in which $\Pr(\mathcal{M}(D) = O)$ represents the density function in the continuous or the probability function in the discrete case.

## 4 Survey of DP Resilience Bounds against Attacks on Private Training Data

In this section, we present an overview of the existing attacks on training databases and the proven resilience bounds obtained when the learning mechanism satisfies DP. We present the flaws and limitations of current attack performance metrics and generic bounds, which motivate the definition and analysis of a consistent general attack performance metric in Section 7.

Adversaries are commonly distinguished in terms of the information available for their attack. The first distinction is whether the adversary has white-box or black-box access to the trained mechanism $\mathcal{M}_D$. A white-box attack has information about the internal structure of $\mathcal{M}_D$ while a black-box attack only knows input-output pairs $(x, \mathcal{M}_D(x))$ derived from querying $\mathcal{M}_D$ [16]. The second distinction is how much information the adversary already has about the target dataset and record. Following the state-of-the-art nomenclature, we classify them as follows:

*Definition 4.1 (Weak Adversary).* Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a learning algorithm and $O = \mathcal{M}_D$ be the output of $\mathcal{M}$ after training with dataset $D$ of size $n$. A *Weak Adversary* has access to

(1) the data distribution $\pi^n$,
(2) any released parameters pertaining to the output trained model $O$ (if white-box) or the pairs $(x, O(x))$ (if black box).

*Definition 4.2 (Informed Adversary [1]).* Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a learning algorithm and $O = \mathcal{M}_D$ be the output of $\mathcal{M}$ after training with dataset $D$ of size $n$. Let $z \in D$ be an arbitrary record and $D_- = D \setminus \{z\}$ denote the remaining records. An *Informed Adversary* has access to

(1) the fixed dataset $D_-$ and the distribution of data records $\pi$,
(2) any released parameters pertaining to the output trained model $O$ and the mechanism $\mathcal{M}$,
(3) (optional) background-knowledge aux about the unknown record $z$.

*Definition 4.3 (Strong Adversary [18]).* Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a learning algorithm and $O = \mathcal{M}_D$ be the output of $\mathcal{M}$ after training with dataset $D$ of size $n$. Let $z \in D$ be an arbitrary record and $D_- = D \setminus \{z\}$ denote the remaining records. A *Strong Adversary* has access to

(1) the fixed dataset $D_-$ and the distribution of data records $\pi$,
(2) any released parameters pertaining to the output trained model $O$ and the mechanism $\mathcal{M}$,
(3) the only two possible values $\{z_0, z_1\}$ for record $z$.

Complementary to the previous classification, we can look at the adversary's goal. In the literature, we find three models of attacks against training data: MIAs [29], AIAs [35], and DRAs [1]. Given a target record, MIAs aim to correctly determine whether this record was part of the training database or not. The AIA seeks to complete the partial information available for the target record that participated in the training database. Depending on the target attribute, this could be a single bit of information (in the case of a binary attribute like positive/negative in a medical test) or several bytes worth of information (for example, the address). The DRA is the most general attack and aims to infer any quantity of information from a record in the dataset. The formalization of this attack can also be applied to MIAs and AIAs, as we will show in Section 8 where we give bounds on the attack resilience of MIAs and AIAs by modeling them as particular cases of DRAs.

### 4.1 Membership Inference Attacks

Among the listed attacks, MIAs have been investigated most concerning the protection offered by DP (e.g., [14, 27, 18, 35]). We find a variety of bounds for the accuracy of a MIA when DP is enforced. Each establishes a direct relation between the DP privacy parameters $(\varepsilon, \delta)$ and the maximum theoretical accuracy of the attack.

The first bound was derived by Yeom et al. [35] for pure DP. For their bound, they assume black-box access to the model and consider a weak adversary (Definition 4.1). They design the membership experiment for a weak adversary to derive their bound.

The experiment proceeds as follows: It receives the distribution $\pi^n$ from which the dataset $D$ is sampled as input. $D$ is the training dataset for the mechanism $\mathcal{M}$ that produces the trained mechanism $\mathcal{M}_D$. The data owner then chooses a binary variable $b$ at random. This variable determines the type of record $z$ that the adversary receives. If $b$ is zero, the adversary receives a member record ($z \in D$), otherwise a record is drawn from the underlying distribution of data points ($z \sim \pi$). Finally, the adversary carries out the attack $A(z, \mathcal{M}_D, \pi)$, which outputs a bit indicating the membership status of $z$. The membership experiment is formalized in Definition 4.4.

*Definition 4.4 (Weak Membership Experiment [35]).* Let A be an attack, $\mathcal{M}$ a learning algorithm and $\pi^n$ a distribution over databases. The membership experiment proceeds as follows:

(1) Sample $D \sim \pi^n$ and let $\mathcal{M}_D = \mathcal{M}(D)$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) If $b = 0$, draw $z \in D$, else draw $z \sim \pi$.
(4) $\text{Exp}^{\text{MIA}}(A, \mathcal{M}, n, \pi) = 1$ if $A(z, \mathcal{M}_D, \pi) = b$, and 0 otherwise.

The success of a MIA is commonly measured with the membership advantage as per Definition 4.5. It expresses how well an adversary can distinguish between a record from the database $z \in D$ and a record sampled from the data distribution $z \sim \pi$ due to the access to the trained model.

*Definition 4.5 (Membership Advantage).* Let A be the attack described in Definition 4.4. The membership advantage for the membership inference attack is defined as

$$\text{Adv}_{\text{MIA}}(A, \mathcal{M}, \pi) = 2\Pr[\text{Exp}^{\text{MIA}}(A, \mathcal{M}, n, \pi) = 1] - 1.$$

Note that if we model our problem as hypothesis testing where $b = 0$ is the null hypothesis ($z$ belongs to the database) and $b = 1$ is the alternative hypothesis ($z$ was drawn from the distribution of possible data records), the membership advantage can be expressed as the true positive rate minus the false positive rate:

$$\begin{aligned}\text{Adv}_{MIA} &= \Pr_{\substack{D\sim\pi^n \\ \theta\sim\mathcal{M}(D)}}\left(\Pr_{Z\sim D}(A(\theta, Z, \pi) = 0) - \Pr_{Z\sim\pi}(A(\theta, Z, \pi) = 0)\right) \\ &= \Pr_{\substack{D\sim\pi^n \\ \theta\sim\mathcal{M}(D)}}(A(\theta) = 0|b = 0) - \Pr_{\substack{D\sim\pi^n \\ \theta\sim\mathcal{M}(D)}}(A(\theta) = 0|b = 1) \quad (1)\end{aligned}$$

*Remark.* An advantage ranges between -1 and 1 since it is a subtraction of probabilities. In particular, the membership advantage does not give the adversary credit for correctly identifying a point drawn from $\pi$ (i.e., $b = 1$), even if it is a member of $D$. As a result, the maximum advantage that an adversary can hope to achieve is $1 - \mu(n, D)$ where $\mu(n, D) = \Pr_{D\in\pi^n, z\sim\pi}[z \in D]$ is the probability of resampling an individual from the training set into the general population [35]. In a weak membership experiment, we expect the set of possible members with size $n$ to be large. Then, $\mu(n, D)$ has almost no effect. However, we will see that the stronger the attack gets, i.e., the lower the number of potential members, the larger $\mu$ becomes. In the extreme case of a strong membership inference attack, it results in $\mu(D, 2) \geq 1/2$.

The success of a weak membership experiment under DP private learning was studied by Yeom et al. who obtained the first bound:

$$\text{Adv}_{\text{MIA}}(A, \mathcal{M}, n, \mathcal{D}) \leq \mathrm{e}^{\varepsilon} - 1$$

Later on, this bound was improved by Erlingsson et al.:

Theorem 4.6 (Upper Bounds for Weak MIAs under Pure DP [14]). *Let $\mathcal{M}$ be an $\varepsilon$-DP mechanism and A an attack. The membership advantage* $\text{Adv}_{\text{MIA}}$ *satisfies*

$$\text{Adv}_{\text{MIA}}(A, \mathcal{M}, n, \mathcal{D}) \leq 1 - \mathrm{e}^{-\varepsilon}$$

Therefore, the probability that the adversary can correctly distinguish between members and non-members of a dataset is limited directly by the privacy budget $\varepsilon$. The higher $\varepsilon$ becomes, the weaker this guarantee is.

Humphries et al. [18] improve the bound given by Erlingsson et al. [14] and introduce a tighter bound under a strong adversary assumption (Definition 4.3). The strong adversary MIA experiment is a particular case of the weak adversary experiment and can be formalized as follows:

*Definition 4.7 (Strong Membership Experiment with Resampling).* Let A be an attack, $\mathcal{M}$ a learning algorithm and $\pi$ a distribution over $\{z_0, z_1\}$. The membership experiment proceeds as follows:

(1) Sample $z' \sim \pi$ and let $\mathcal{M}_{z'} = \mathcal{M}(D \cup \{z'\})$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) If $b = 0$, draw $z = z'$, else draw $z \sim \pi$.
(4) $\text{Exp}^{\text{MIA}}(A, \mathcal{M}, n, \pi) = 1$ if $A(z, \mathcal{M}_{z'}, \pi) = b$, and 0 otherwise.

Note that the reason for allowing resampling $z \sim \pi$ when $b = 1$ is to measure the degree to which $\mathcal{M}_D$ reveals membership and not the effect of any prior background knowledge of $D$ encoded in $\pi$ [35].

Therefore, for a strong adversary under uniform priors, we can simplify the experiment as proposed in [18]:

*Definition 4.8 (Simplified Strong Membership Experiment [18]).* Let $A^s$ be an attack, $\mathcal{M}$ a learning algorithm and $\pi^n$ a distribution over databases. The membership experiment proceeds as follows:

(1) $A^s$ receives $D_-$, $\{z_0, z_1\}$ and $\pi$.
(2) Choose $\alpha \leftarrow \{0, 1\}$ uniformly at random.
(3) Train $O = \mathcal{M}_{D_{z_\alpha}}$.
(4) $\text{Exp}_s^{MIA}(A, \mathcal{M}, \pi) = 1$ if $A^s(\mathcal{M}_{D_{z_\alpha}}) = \alpha$, and 0 otherwise.

Applying Definition 4.12 to this stronger experiment, we get the advantage of a strong membership experiment:

$$\text{Adv}^s_{\text{MIA}} = 2\Pr[\text{Exp}^{\text{MIA}}_s] - 1 \quad (2)$$

Note that, as we prove in Appendix B, we have the following relationship between the advantage of a strong membership experiment with resampling (Definition 4.4) and without resampling (Definition 4.7):

$$\text{Adv}_{\text{MIA}} = 2\Pr[\text{Exp}^{MIA}] - 1 = \Pr[\text{Exp}^{MIA}_s] - \frac{1}{2} = \frac{1}{2}\text{Adv}^s_{MIA}$$

This is coherent with the fact that $\text{Adv}_{\text{MIA}}$ is upper-bounded by $\frac{1}{2}$ in a strong membership experiment with resampling (cf. Section 4.1). $\text{Adv}^s_{\text{MIA}}$, on the other hand, is upper-bounded by 1 when the adversary always identifies the membership correctly and lower-bounded by 0 when the adversary decides randomly. We call $\text{Adv}^s_{\text{MIA}}$ the strong membership advantage.

Humphries et al. [18] prove a bound of the strong membership advantage that holds for all $\varepsilon \geq 0$ and $0 \leq \delta \leq 1$ [18]. This bound is stated in Theorem 4.9.

Theorem 4.9 (Upper Bound for MIAs under $(\varepsilon, \delta)$-DP [18]). *Let $\mathcal{M}$ be an $(\varepsilon, \delta)$-DP mechanism and A an attack carried out by a strong adversary. Then, the membership advantage* $\text{Adv}_{MIA}$ *satisfies*

$$\text{Adv}^S_{MIA}(A, \mathcal{M}, n, \mathcal{D}) \leq \frac{\mathrm{e}^{\varepsilon} - 1 + 2\delta}{\mathrm{e}^{\varepsilon} + 1}.$$

Note that Humphries et al.'s bound is tighter than the previously discussed bounds for $\delta = 0$. Moreover, they prove that their bound also holds for the weakest adversary where the joint distribution of a sequence of $n$ members and one non-member does not depend on the order of the sequence.

## 4.2 Attribute Inference Attacks

In the case of AIAs, instead of inferring the membership status of a fully known record $z$, the adversary aims to complete the information they have available on $z$ by inferring a target attribute of this record [1]. Formally, the data point is now represented as a triple $z = (v, t, y)$ where $(v, t) = x \in \mathcal{X}$. $v$ describes the known features, $t$ the target attribute, and $y \in \mathcal{Y}$ the output under mechanism $\mathcal{M}_D$. A fixed function $\phi$ with domain $\mathcal{Z}$ describes the knowledge of the adversary about a given record, $\phi(z) = v$. The function $\varphi(z)$ describes the correct value of the target attribute $t$. The attribute experiment is formalized in Definition 4.10.

*Definition 4.10 (Weak Attribute Experiment [35]).* Let $A$ be an attack, $n$ a positive integer and $\pi^n$ a distribution over databases. The attribute experiment proceeds as follows:

(1) Sample $D \sim \pi^n$.
(2) Choose $b \leftarrow \{0, 1\}$ uniformly at random.
(3) If $b = 0$, draw $z \in D$, else draw $z \sim \pi$.
(4) $\text{Exp}^{\text{AIA}}(A, \phi(z), \mathcal{M}, \pi) = 1$ if $A(\phi(z), \mathcal{M}_D, \pi) = \varphi(z)$, and 0 otherwise.

The success of an AIA is also measured by an advantage, defined in Definition 4.11. It measures the amount of information about the target attribute $t$ leaked by mechanism $\mathcal{M}$. The attribute advantage compares the probability that the adversary successfully identifies $t$ when $z$ is a record from the dataset $D$ with the probability of correctly identifying $t$ for a record drawn from the underlying distribution $\pi$. For instance, if the adversary generally has a high chance of guessing the attribute correctly, the attribute advantage will be lower because the probability of correctly inferring the value of the target attribute for a record from the underlying distribution increases. This is done because we are interested specifically in the information leaked by $\mathcal{M}$.

*Definition 4.11 (Attribute Advantage [35]).* The attribute advantage of $A$ is defined as

$$\text{Adv}_{\text{AIA}}(A, \mathcal{M}, \pi) =$$
$$\Pr[\text{Exp}^{\text{AIA}}(A, \mathcal{M}, \pi) = 1 | b = 0] - \Pr[\text{Exp}^{\text{AIA}}(A, \mathcal{M}, \pi) = 1 | b = 1].$$

The previous section established that DP protects against MIAs for certain values of $\varepsilon$ and $\delta$. However, the only formal adversarial bound for AIAs that has been established applies when this attack is performed using a MIA as a baseline. This is not the only strategy to run an AIA but a particular possibility.

In essence, the attack tries all possible values $t_i$ for the target attribute $t$ and uses them to complete a candidate record $z_i = (v, t_i, y)$. This candidate record $z_i$ is then the input to the MIA $A_{\text{MIA}}(z_i, \mathcal{M}, n, \pi)$. If the record can be identified as a member of dataset $D$, the adversary can infer that $t_i$ was the correct value for the target attribute. This strategy is formalized in Definition 4.12.

*Definition 4.12 (AIA Strategy via MIA [35]).* Let $t_1, \ldots, t_m$ be the possible values of target $t$. The attack $A_{\text{AIA}\to\text{MIA}}$ has access to a membership inference attack $A_{\text{MIA}}$. With the input $\phi(z)$, $\mathcal{M}$, $n$ and $\pi$, the attack proceeds as follows:

(1) Choose $t_i$ uniformly at random from $\{t_1, \ldots, t_m\}$.
(2) Let $z' = \phi^{-1}(\phi(z))$ be a candidate record. Change the target attribute $t$ such that $\varphi(z') = t_i$.
(3) Run the MIA to obtain $b \leftarrow A_{\text{MIA}}(z', \mathcal{M}, n, \mathcal{D})$.
(4) If $b = 0$, output $t_i$.

One limitation of this strategy is immediately obvious: the attack performance of this variant of the AIA is at most equal to the performance of the MIA. The following discussion elaborates precisely how much worse it is.

When comparing the advantage of attack $A_{\text{AIA}\to\text{MIA}}$ to the performance of $A_{\text{MIA}}$, it can be shown that the attribute advantage of $A_{\text{AIA}\to\text{MIA}}$ lies within a constant factor of the membership advantage of $A_{\text{MIA}}$. This relationship is given in Theorem 4.13.

Theorem 4.13 (Attribute Advantage Bound by Membership Advantage [35]). *Let $A_{\text{AIA}\to\text{MIA}}$ be the attack described in Definition 4.12 which uses $A_{\text{MIA}}$. For the advantage, it holds that*

$$\text{Adv}_{\text{AIA}}(A_{\text{AIA}\to\text{MIA}}, \mathcal{M}, n, \mathcal{D}) = \frac{1}{m}\text{Adv}_{\text{MIA}}(A_{\text{MIA}}, \mathcal{M}, n, \mathcal{D})$$

Without background knowledge, the probability of correctly guessing the target attribute is $\frac{1}{m}$, where $m$ is the number of possible attribute values.

Since an AIA using an MIA can only perform $\frac{1}{m}$-times as well as the MIA performs, the performance for this attack is deficient for attributes with a high number of possibilities, e.g., last names or addresses.

## 4.3 Data Reconstruction Attacks

MIAs and AIAs are particular cases of the more general DRAs [1]. The adversary in a DRA aims to produce an accurate reconstruction of the target record $z$. This reconstruction can aim for any information about the record $z$, for instance, various attributes or its membership status. As an example, for an image database, we could try to guess the target's eye color (AIA) or reconstruct an image as close as possible to the individual's record. This attack is formalized as follows:

*Definition 4.14 (DRA Experiment with Informed Adversary [1]).* Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a mechanism, $D$ a dataset that contains target record $z$, $l$ an error-function and aux the background-knowledge. Additionally, let $D = D_- \cup \{z\}$. Then, the reconstruction attack $A$ proceeds as follows

(1) $z' \leftarrow A(\mathcal{M}, D_-; \text{aux})$.
(2) Output $l(z, z')$ .

The attack produces a candidate record $z'$ and returns a measure of the success of the attack based on the reconstruction error $l$. A lower error indicates a more successful reconstruction. Various error metrics are possible for $l$, which capture different aspects of information leakage. Balle et al. use the mean square error, the Kullback–Leibler (KL) divergence, and image-specific metrics such as the Learned Perceptual Image Patch Similarity (LPIPS) in their analysis [1].

Balle et al. [1] propose a metric of accuracy for this type of attack: Reconstruction robustness (ReRo) as stated in Definition 4.15.

*Definition 4.15 ($(\eta, \gamma)$-Reconstruction Robustness (ReRo) [1] ).* Let $\pi$ be a prior over $\mathcal{Z}$ and $l : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function. Mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is $(\eta, \gamma)$-reconstruction robust with respect to $\pi, l$ if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction attack

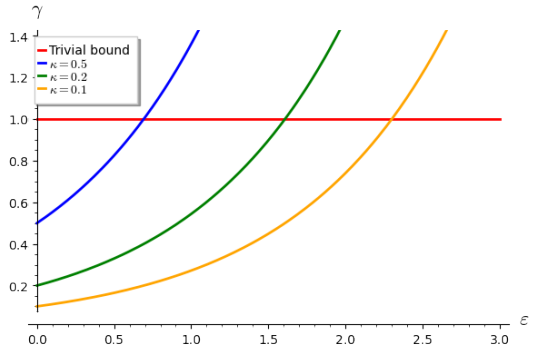



**Figure 1: Bound on $\gamma$ for $(\eta, \gamma)$-ReRo proven by Balle et al. for different $\kappa$ values. The red line indicates the value of 1. Since $\gamma$ represents a probability, a bound larger than one trivially is not tight. This figure shows how the bound obtained by Proposition 4.16 exceeds one in most cases, even for small values of epsilon.**

$A\colon \Theta \to \mathcal{Z}$ it holds that

$$\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_{-}\cup\{Z\})}}[l(Z, A(\theta)) \leq \eta] \leq \gamma. \tag{3}$$

This definition prevents the reconstruction of unknown target $z$ from attaining an error lower than $\eta$ with a probability larger than $\gamma$. The parameter $\eta$ defines what we consider a successful attack, while $\gamma$ bounds the probability of success. Therefore, the larger the $\eta$ and the lower the $\gamma$, the less effective the attack.

Furthermore, Balle et al. [1] demonstrate that DP mechanisms imply ReRo, i.e., protect against reconstruction attacks. The degree of protection is then parameterized by the DP privacy parameters $\varepsilon$ and the baseline error $\kappa$ that is defined as

$$\kappa_{\pi,l}(\eta) = \sup_{z'\in\mathcal{Z}} \Pr_{Z\sim\pi}[l(Z, z') \leq \eta]. \tag{4}$$

Intuitively, $\kappa$ describes the probability of an oblivious attack to succeed. "Oblivious" means that the adversary does not use any information from the training model or the trained data but randomly tries to guess a correct reconstruction [1]. Therefore, the success of an oblivious adversary is the same whether or not the output of the trained model is shared or whether the target participated in the training.

Proposition 4.16 ($\varepsilon$-DP implies ReRo [1]). *Let $\pi, l$ and $\eta > 0$ and $\kappa = \kappa_{\pi,l}(\eta)$. If a mechanism $\mathcal{M}$ satisfies $\varepsilon$-DP, then it also satisfies $(\eta, \gamma)$-ReRo with $\gamma = \kappa e^{\varepsilon}$.*

Figure 1 shows how the bound on $\gamma$ varies for different values of $\varepsilon$ and $\kappa$ using Proposition 4.16.

ReRo is a performance metric that applies for any attack against training data, and Proposition 4.16 allows us to relate this performance metric to the DP parameters. Furthermore, it is the first metric that allows us to measure the success of an attack that does not perform a perfect reconstruction of the target information. Since MIAs, AIAs, and DRAs can be modeled jointly by considering the informed adversary (as presented in Definition 4.2), ReRo is a promising performance metric that, together with the bound from Proposition 4.16, could unify the state of the art and enhance the interpretability of DP. However, we point out two main problems.

First, we can see in Figure 1 that the bound of Proposition 4.16 is not tight since it yields values over 1 when bounding a probability where 1 is the trivial bound. The untightness of Proposition 4.16 is a problem since using this bound as a criterion for choosing our privacy parameters would result in a poor utility of the results due to an overestimation of privacy leakage. This motivates us to obtain a tighter bound on ReRo under DP protection that we present in Section 5.

Second, ReRo does not have the same interpretation as the state-of-the-art performance metrics for MIAs and AIAs. Both metrics, $\text{Adv}_{\text{MIA}}$ and $\text{Adv}_{\text{AIA}}$, measure the amount of information (either about the membership or the attributes) of a target that is leaked specifically by participating in the training of the mechanism $\mathcal{M}$. To this end, both advantages include a correction factor to remove the probability of correctly guessing information about a target even if the target did not participate in the training database. This avoids overestimating the privacy risk if the attack's success is based only on the learned properties of the population and not on the individuals in the training database.

However, ReRo does not contain any correcting factor, and therefore, its interpretability changes with respect to the usual metrics since we are overestimating the leakage because of the *statistical privacy fallacy* phenomenon [4]. In section Section 6, we elaborate on these problems, and in Section 7, we propose the new performance metric Unbiased Reconstruction Robustness (U-ReRo) that takes the statistical privacy fallacy into account and generalizes the state-of-the-art attack performance metrics.

*Attack Summary:* In this section, we discussed state-of-the-art bounds relating to DP and the performance of three important attacks on private data. An overview of these attacks can be found in Table 2. We can see that, even if we find bounds for all the attacks, these bounds are not general or comparable.

The first bound for the AIA [35] only holds for adversaries that use a MIA to accomplish an AIA. This is a strong limitation in the choice of attack strategy. Another problem is that each bound was derived for the specific performance metric with different interpretations. For instance, the bound for MIAs bounds the membership advantage, whereas the bound for DRAs bounds the probability of an accurate reconstruction. Since they represent different measures, we cannot compare which attack is mitigated better by DP.

On the other hand, the bound for the ReRo of DRAs using Proposition 4.16 applies to all attacks and allows us to establish a general bound. However, this bound is not tight and does not consider the bias of the privacy fallacy [4] as we will see in Section 6.

To solve these problems, we propose the standardized performance metric U-ReRo in Section 7, using the DRA model as a baseline and analyzing the other attacks as particular scenarios of this attack model. This allows for a general and consistent comparison of attack performance under DP.

## 5 Tighter bound for ReRo

Proposition 4.16 proves that DP bounds the reconstruction robustness of reconstruction attacks. However, it does not provide a tight

| Attack | Assumptions | Performance Metric | SOTA Bound | Our Improved Bound |
|---|---|---|---|---|
| MIA [18] | Strongest | $\mathrm{Adv}^s_{MIA}$ (Eq. 2) $\equiv (0, \frac{\gamma}{2})$-U-ReRo | $\frac{\gamma}{2} \equiv \mathrm{Adv}^s_{\mathrm{MIA}} \leq \frac{e^\varepsilon-1}{e^\varepsilon+1}$ | - |
| MIA | Strongest | $(0,\gamma)$-ReRo (Def. 4.15) | $\gamma \leq \frac{1}{2}e^\varepsilon$ | $\gamma \leq \frac{1}{2}\left(\frac{e^\varepsilon-1}{e^\varepsilon+1}+1\right)$ |
| MIA | Informed<br>Uniform prior<br>$m$ possibilities | $\mathrm{Adv}_{MIA}$(Def. 4.11)$\equiv (0,\gamma)$-U-ReRo | $\mathrm{Adv}_{MIA} \leq \frac{e^\varepsilon-1}{e^\varepsilon+1}$ | $\gamma \equiv \mathrm{Adv}_{AIA} \leq \min\{\frac{1}{m}(e^\varepsilon-1), \frac{m-1}{m}\frac{e^\varepsilon-1}{e^\varepsilon+1}\}$ |
| AIA [35] | Strategy via MIA | $\mathrm{Adv}_{AIA}$(Def. 4.11) | $\mathrm{Adv}_{\mathrm{AIA}} = \frac{1}{m}\mathrm{Adv}_{\mathrm{MIA}}$ | - |
| AIA | Informed | $(0,\gamma)$-ReRo (Def. 4.15) | $\gamma \leq \frac{1}{m}e^\varepsilon$ | $\gamma \leq \min\{\frac{e^\varepsilon}{m}, \frac{m-1}{m}\left(\frac{e^\varepsilon-1}{e^\varepsilon+1}+1\right)\}$ |
| AIA | Informed<br>Uniform prior<br>$m$ possibilities | $\mathrm{Adv}_{AIA}$(Def. 4.11))$\equiv (0,\gamma)$-U-ReRo | - | $\gamma \equiv \mathrm{Adv}_{AIA} \leq \min\{\frac{1}{m}(e^\varepsilon-1), \frac{m-1}{m}\frac{e^\varepsilon-1}{e^\varepsilon+1}\}$ |
| DRA [1] | Informed | $(\eta,\gamma)$-ReRo (Def. 4.15) | $\gamma \leq \kappa_\eta e^\varepsilon$ | - |
| DRA | Informed | $(0,\gamma)$-ReRo (Def. 4.15) | $\gamma \leq \kappa_0 e^\varepsilon$ | $\gamma \leq \min\{\kappa_0(e^\varepsilon), \kappa_0(m-1)\left(\frac{e^\varepsilon-1}{e^\varepsilon+1}+1\right)\}$ |
| DRA | Informed | $(\eta,\gamma)$-U-ReRo (Def. 7.1) | - | $\gamma \leq \min\{\kappa_\eta(e^\varepsilon-1), \frac{e^\varepsilon-1}{e^\varepsilon+1}\}$ |
| DRA | Informed | $(0,\gamma)$-U-ReRo (Def. 7.1) | - | $\gamma \leq \min\{\kappa_0(e^\varepsilon-1), \kappa_0(m-1)\frac{e^\varepsilon-1}{e^\varepsilon+1}+\kappa_0-\kappa_0^-\}$ |

**Table 2: Overview of Attacks and Bounds for $\varepsilon$-DP Mechanisms.**

bound, i.e., the protection that DP offers against DRAs is underestimated by this result.

This is indicated by two observations: First, we see in Figure 1 and Figure 2 that for $\kappa = 0.5$ and $\varepsilon \geq 1$ the bound of ReRo is larger than one, which is a trivial bound given that ReRo is a probability. The second hint is given by Balle et al. [1]. According to the authors, no meaningful protection can be ensured in theory when $\varepsilon \geq 1$, while empirically, even large values show a decrease in the attack performance.

In this section, we give a tighter bound for ReRo for the particular case of perfect reconstruction, $l(z, A(\theta)) = \eta = 0$, which means that $A(\theta)$ perfectly reconstructs the real value $z$. We illustrate the improvement with respect to Balle et al.'s bound in Figure 2.

Proposition 5.1 (Improved Bound for ReRo against perfect reconstruction). *Let $\pi$ be a prior over $\mathcal{Z}$, $l : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function, $|\mathcal{Z}| = m$ and $\kappa_0 = \kappa_{\pi,l}(0)$. If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ satisfies $\varepsilon$-DP, then it also satisfies $(0,\gamma)$-ReRo with $\gamma = \min\{\kappa_0 e^\varepsilon, \kappa_0\left(1 + (m-1)\frac{e^\varepsilon-1}{e^\varepsilon+1}\right)\}$, i.e.,*

$$\Pr_{\substack{z\sim\pi \\ \theta\sim\mathcal{M}(D_z)}}[l(z, A(\theta)) = 0] \leq \min\{\kappa_0 e^\varepsilon, \kappa_0\left(1 + (m-1)\frac{e^\varepsilon-1}{e^\varepsilon+1}\right)\}$$

We can simplify the bound in the case of uniform prior probability, where all possible reconstructions are equally likely before performing the attack. We denote a set of possible reconstructions

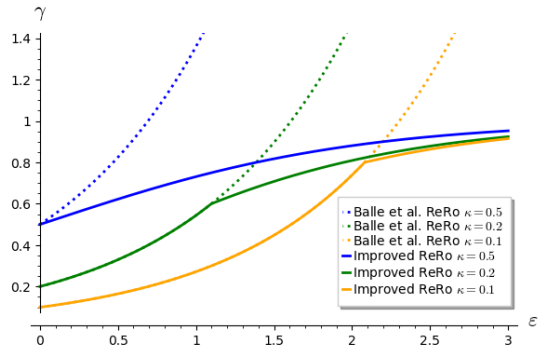



**Figure 2: Comparison of the ReRo bound for a perfect reconstruction from [1] with our bound from Proposition 5.1 for uniform priors. The dashed lines correspond to the bound for ReRo using Proposition 4.16 (Balle et al.) under $\varepsilon$-DP. The continuous lines correspond to the improved bound that we prove in Proposition 5.1.**

of size $m$ as $\pi = U[m]$. Then, we can simplify the bound to:

$$\Pr_{\substack{z\sim U[m]\\ \theta\sim\mathcal{M}(D_z)}}[l(z,A(\theta))=0] \le \min\{\frac{e^\varepsilon}{m}, \frac{1}{m}+\frac{m-1}{m}\left(\frac{e^\varepsilon-1}{e^\varepsilon+1}\right)\}$$

As we see in Figure 2, this result provides a significantly better bound when the set of possible reconstructions is small. Moreover, while with Proposition 4.16 no protection is formally guaranteed for $\varepsilon \approx 1$ and $\kappa = 0.5$, our bound never surpasses 1.

## 6 ReRo and the Privacy Fallacy

The *Statistical Inference Privacy Fallacy* [4] can be illustrated with the classic example of the smoking database [21]. Imagine a model trained on medical data from a database $D$ that learns that the probability of having cancer is high (e.g., 0.9) for smokers. An adversary who knows that an individual $z \in D$ smokes directly infers that $z$ has cancer. However, this is not an issue that could be prevented by privacy mechanisms: Even if $z$ would deny volunteering their data for the training process, an adversary could still infer that they have cancer. The reason is that the correlation between the two attributes is a global statistic learned from the population without ever exposing $z$'s private information directly.

The definition of ReRo as an attack performance metric overestimates the privacy leakage of the learning mechanism due to this fallacy. ReRo calculates the probability that the attack successfully reconstructs a database member. However, it does not distinguish whether the successful reconstruction is due to the information $\mathcal{M}$ learned about $z$ (an actual privacy leakage) or some other source. Examples include global statistics about the distribution of records $\pi$ learned during training or an auxiliary source such as previous studies. If the success of the attack comes from an external source, such as a study showing that smoking causes cancer or the ability to learn the correlation between smokers and cancer by $\mathcal{M}$, then there is no privacy risk for $z$ because the consequences would be the same even without participating. To illustrate, we give the following example. Given a data universe $\pi^n$ from which two databases with disjoint sets of individuals $D_1, D_2$ are drawn, we use $D_1$ to train $\mathcal{M}$ as a classifier that, given a set of medical conditions $x$, outputs whether this patient has cancer. Now, we consider a reconstruction attack that tries to infer whether $z$ has cancer using the background knowledge $x$. Since $D_1, D_2$ are identically distributed, we can use $\mathcal{M}_{D_1}$ to classify whether a set of medical conditions implies cancer and perform an attribute inference attack on $D_2$. In this case, for all $z_2 \in D_2$, $\Pr_{\theta\sim\mathcal{M}(D_1)}(l(z_2,A(\theta)) \le \eta)$ will be as large as the utility of our classifier. However, no privacy leakage can occur for $z_2 \in D_2$ since this individual never volunteered data for the training. Therefore, if $\Pr_{\theta\sim\mathcal{M}(D_{z_1})}[l(z_1,A(\theta)) \le \eta] = \Pr_{\theta\sim\mathcal{M}(D_{z_1})}[l(z_2,A(\theta)) \le \eta]$ with $z_1 \in D_1$ and $z_2 \in D_2$, the successful reconstruction of $z_1$ is not a privacy issue since it is the same for records that did not participate, such as $z_2$. The information leading to the successful inference is external to the private information shared by $z_1$.

More formally, if

$$\gamma = \Pr_{\substack{Z\in\pi\\ \theta\sim\mathcal{M}(D_Z)}}[l(Z,A(\theta)) \le \eta] = \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}} \Pr_{Z'\notin D_Z}[l(Z',A(\theta)) \le \gamma],$$

we know that the reconstruction success does not represent a privacy risk even if the $\gamma$ in $(\gamma,\eta)$-ReRo is arbitrarily large. This relies on the probability of a correct reconstruction for participants being identical to the one of a correct reconstruction for non-participants.

Both membership advantage and attribute advantage incorporate a correction factor to eliminate the bias of a successful attack based on the privacy fallacy rather than the privacy leakage. However, we demonstrated that ReRo cannot distinguish between both cases. For the $\mathrm{Adv}_{MIA}$, the correcting factor $\Pr(A=0|b=1)$ is removed from the probability of correctly guessing a member. Analogously, $\mathrm{Adv}_{AIA}$ includes the correction factor $\Pr(Exp^{AIA}(A,\mathcal{M},\pi)=1|b=1)$. It adjusts the advantage by subtracting the probability of successful attribute reconstruction when the target was drawn from the overall data distribution instead of the actual database. This correction term ensures that only the privacy leakage from training $\mathcal{M}$ is measured, accurately reflecting the impact on the participant's privacy and avoiding overestimating the risk due to the privacy fallacy.

This mismatch between the advantages and ReRo is especially tangible if we compute the ReRo bound for the perfect reconstruction of membership. In this case, we see that even if both metrics are measuring the same attack resilience, they yield different values:

Proposition 6.1. *Given a learning mechanism $\mathcal{M}$, consider a strong membership experiment (Def. 4.8) where $D_-$ is known and the two possible members $\{z_0, z_1\}$ are uniformly distributed. Given $l$, the membership error function (Eq. 7), such that $l(z,z')=0$ if $z=z'$ and 1 otherwise, we obtain:*

$$\mathrm{Adv}^s_{\mathrm{MIA}} = 2 \Pr_{\substack{z_i\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_i})}}[l(z_i,A(\theta))=0] - 1.$$

As a consequence, if $\mathcal{M}$ satisfies $(0,\gamma)$-ReRo, it satisfies $\mathrm{Adv}^s_{\mathrm{MIA}} = 2\gamma - 1$. Since both metrics output different values for the same scenario, it is clear that they are not measuring the same information. Since $\gamma \in [0,1]$, we derive from this formula that $\gamma$ is always bigger than the membership $\mathrm{Adv}^s_{\mathrm{MIA}}$, showing that $\gamma$ is an overestimation of the actual privacy risk of a MIA.

## 7 Unbiased Reconstruction Robustness (U-ReRo)

To address the fact that ReRo does not account for the privacy fallacy, we propose a new general performance metric: U-ReRo. It can be applied to DRAs, AIAs, and MIAs and takes the privacy fallacy into account. Furthermore, we prove a result that relates U-ReRo with DP's privacy parameters and apply it to the particular cases of MIAs and AIAs.

*Definition 7.1 (Unbiased Reconstruction Robustness (U-ReRo)).* Let $\pi$ be a prior over $\mathcal{Z}$ and $l\colon \mathcal{Z}\times\mathcal{Z}\to\mathbb{R}_{\ge 0}$ a reconstruction error function. A randomized learning mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ is $(\eta,\gamma)$-unbiased reconstruction robust,$(\eta,\gamma)$-U-ReRo, with respect to $\pi$ and $l$ if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction attack $A\colon \Theta\to\mathcal{Z}$ we have

$$\Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}}[l(Z,A(\theta)) \le \eta] - \mathbb{E}_{Z_0\sim\pi}\left[\Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_{Z_0})}} l(Z,A(\theta)) \le \eta\right] \le \gamma. \tag{5}$$

While ReRo bounds a probability, U-ReRo bounds an advantage. It compares the probability that the adversary correctly reconstructs the target record when it is a record from the dataset $D$ with the probability of correctly reconstructing a record drawn from the underlying distribution $\pi$. In this way, U-ReRo only gives the adversary credit for the probability of correctly reconstructing the target record using exclusively the information that the model has learned about that record. The second term of Equation (5) corresponds to the probability of correctly reconstructing (with a reconstruction error bounded by $\eta$) any record $z$, independently if $z$ was part of the training database or not. This correcting factor removes those cases in which the adversary would correctly reconstruct the target value even when the target did not participate in the database. Therefore, it gives us a more sensible measure of the privacy risk of participation.

The benefit of using U-ReRo instead of ReRo is exemplified in Figure 3. We plot the ReRo bound for $k = 0.5$ in red and the U-ReRo bound for the same $k$ value in blue. We can see that U-ReRo provides a significantly smaller risk value due to the correction factor. By using our metric, practitioners can choose larger $\varepsilon$ values, which are associated with less noise in DP mechanisms and hence yield more utility for the same privacy risk.

Furthermore, we provide a bound for U-ReRo under DP protection that can be used in practice to adjust the noise in a learning mechanism to a desired limitation of attack performance:

Theorem 7.2 ($\varepsilon$-DP Implies $(\eta, \gamma)$-U-ReRo). *Let $\pi, l$ and $\eta \geq 0$ follow Definition 7.1, and $\kappa_\eta \equiv \kappa_{\pi,l}(\eta)$. If a mechanism $\mathcal{M}$ satisfies $\varepsilon$-DP, then it also satisfies $(\eta, \gamma)$-U-ReRo with*

$$\gamma = \min\{\kappa_\eta(e^\varepsilon - 1), \frac{e^\varepsilon - 1}{e^\varepsilon + 1}\}$$

Note that this bound applies to any error $\eta$ of our reconstruction. However, for $l(z, A(\theta)) = \eta = 0$, which means that $A(\theta)$ perfectly reconstructs the real value $z$, we can obtain a tighter result.

Theorem 7.3 ($\varepsilon$-DP Implies $(0, \gamma)$-U-ReRo). *Let $\pi$ and $l$ follow Definition 7.1, $|\mathcal{Z}| = m$ and $\kappa_0 = \kappa_{\pi,l}(0)$. If a mechanism $\mathcal{M}$ satisfies $\varepsilon$-DP, then it also satisfies $(0, \gamma)$-ReRo with*

$$\gamma = \min\{\kappa_0(e^\varepsilon - 1), \kappa_0(m-1)\frac{e^\varepsilon - 1}{e^\varepsilon + 1} + \kappa_0 - \kappa_0^-\},$$

*where, $\kappa_0^- \equiv \kappa_{\pi,l}^-(0) := \inf_{z' \in \mathcal{Z}} \Pr_{z\sim\pi}[l(z, z') = 0]$ is the lower baseline error.*

If we consider $\pi$ uniform over a set of possible reconstructions of size $m$, we obtain:

$$\gamma = \min\{\frac{1}{m}(e^\varepsilon - 1), \frac{m-1}{m}\frac{e^\varepsilon - 1}{e^\varepsilon + 1}\} \tag{6}$$

We show in Figure 3 how this bound of $\gamma$ increases with $\varepsilon$ for different sizes of the possible reconstruction set $\mathcal{Z}$. We also show the upper bound in gray, independent of $\kappa$ and $\mathcal{Z}$.

In the following section, we show that U-ReRo is a consistent generalization of the state-of-the-art performance measurements for AIAs and MIAs. That is, by considering these particular scenarios, we obtain the same values. Consequently, we apply our bound proved in Theorem 7.2 to both attacks, obtaining a systematization of attack resilience provided by DP that we summarize in Table 2.

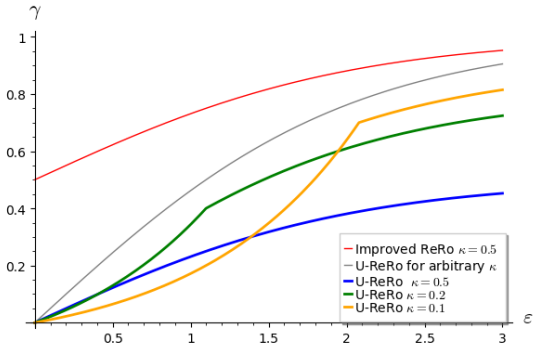



**Figure 3: U-ReRo bounds for a perfect reconstruction from our Theorem 7.3 for different $\kappa$ values under uniform priors. The gray line corresponds to the bound for arbitrary error $\eta$ under arbitrary $\kappa$ using Theorem 7.2. We highlight in red the corresponding bound for ReRo that we proved in Proposition 5.1 to show the extensive over-estimation.**

## 8 Systematization of Attack Bounds via Unbiased Reconstruction Robustness

In this section, we prove that U-ReRo provides equivalent results for MIAs and AIAs to the previously studied performance metrics, namely $\text{Adv}^s_{\text{MIA}}$, $\text{Adv}_{\text{MIA}}$ and $\text{Adv}_{\text{AIA}}$. We demonstrate that it is a correct generalization of attack performance metrics. From this, we derive standardized bounds for all attacks under DP protection using our bound theorem 7.2. In particular, we are the first to give a general bound for the advantage of an attribute inference attack.

### 8.1 Membership Unbiased Reconstruction Robustness under DP

In this section, we prove the equivalence between U-ReRo and the membership advantages. We derive a membership reconstruction bound for the Strongest Adversary in a MIA (Definition 4.3) using our Theorem 7.2. First, we define the following error function $l$ for this scenario:

$$l(z, z') = \begin{cases} 0 & \text{if } z = z', \\ 1 & \text{if } z \neq z', \end{cases} \tag{7}$$

This is a *perfect reconstruction* error function in which the adversary either guesses the member correctly or loses. We can then prove the following relation between the membership advantages and U-ReRo:

Proposition 8.1 ($\text{Adv}_{\text{MIA}} \Leftrightarrow$ U-ReRo). *Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a mechanism. For $A$ an informed MIA, it holds that*

$$\mathcal{M} \text{ is } (0, \gamma)\text{-U-ReRo} \iff \text{Adv}_{\text{MIA}}(A, \mathcal{M}, \pi^n) \leq \gamma,$$

*Additionally, if $A^s$ is a strong MIA under uniform priors (Def. 4.7), then*

$$\mathcal{M} \text{ is } (0, \frac{\gamma}{2})\text{-U-ReRo} \iff \text{Adv}^s_{\text{MIA}}(A, \mathcal{M}, \pi^n) \leq \gamma.$$

This result shows that U-ReRo is equivalent to both membership advantages when applied to the corresponding scenarios. Therefore, our bound for U-ReRo directly bounds $Adv_{MIA}$ and $Adv^s_{MIA}$ as well,

which was not the case for ReRo. The consistency of U-ReRo makes it a suitable general metric for attack performance.

Finally, we apply Theorem 7.3 to obtain a bound for the U-ReRo of a MIA. First, we need to compute $\kappa$. Without any further knowledge about our reconstruction candidates $\{z_1, z_2, \ldots, z_m\}$, i.e., assuming uniform priors, and with the error function from Equation (7), this yields $\Pr_{Z\sim\pi}[l(Z, z') = 0] = \pi(z') = \frac{1}{m}$

The supremum and infimum of these probabilities is $\frac{1}{m}$, thus it follows that $\kappa = \kappa^- = \frac{1}{m}$ under the given assumptions. According to Theorem 7.2, we arrive at the following bound for the probability of accurate reconstruction:

$$\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_z)}}[l(Z, A(\theta)) \leq 0] \leq \min\{\frac{1}{m}(\mathrm{e}^{\varepsilon} - 1), \frac{m-1}{m}\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\}$$

In the case of the strongest MIA, this bound translates into:

$$\Pr_{\substack{Z\sim U\{0,1\} \\ \theta\sim\mathcal{M}(D_z)}}[l(Z, A(\theta)) \leq 0] \leq \frac{1}{2}\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}$$

where $U\{0, 1\}$ denotes the uniform distribution over $\{0, 1\}$.

Therefore, we conclude that $\mathcal{M}$ is at least $(0, \frac{1}{2}\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1})$-U-ReRo against strong MIAs.

## 8.2 Attribute Unbiased Reconstruction Robustness under DP

In this section, we derive the relationship between U-ReRo for an informed adversary in an AIA (Definition 4.2) and the attribute advantage. Using this relation and Theorem 7.2, we give the first bound for the attribute advantage that is independent of the attack strategy.

Analogously to the previous section, we consider only perfect reconstruction. Given the known attributes $\phi(z) = v$ and the target attribute $\varphi(z) = t$, $l(A(\phi(Z), \theta), \varphi(Z)) = 0$ if $A(\theta) = \varphi(z)$ and 1 otherwise. Since $\phi(z)$ is known, it is equivalent to write $l(z, A(\theta)) = 0$ if $z = A(\theta)$ instead. In this scenario, we arrive at the following result:

Proposition 8.2 ($\mathrm{Adv}_{\mathrm{AIA}} \Leftrightarrow$ U-ReRo). *Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a mechanism. For all data distributions $\pi^n$ and for all informed AIA attacks $A$ that know $D_-$ and try to guess the attribute of a target record $z$ from which $\phi(z)$ is known, we have*

$$\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \iff \mathrm{Adv}_{\mathrm{AIA}}(A, \mathcal{M}, \pi^n) \leq \gamma \text{ for all } A.$$

This theorem shows that U-ReRo is equivalent to $Adv_{AIA}$ for the AIA. This is not the case with standard ReRo.

Finally, we use Theorem 7.3 to compute the U-ReRo bound for AIAs. Considering Proposition 8.2, we can use the U-ReRo bound to give, for the first time, a bound for the attribute advantage for arbitrary attack strategies.

To compute the U-ReRo bound, we need to compute the baseline error $\kappa$. Given $\mathcal{Z}$, the set of possible reconstructions, where for all $z, z' \in \mathcal{Z}$ $\phi(z) = \phi(z')$ is the known information about the target entry. The number of possible reconstructions for $z$ depends on the number $m$ of values the target attribute $t$ can take. Let $z_i$ be the reconstruction candidate $z$ where the value of the target attribute $t$ has been replaced with $t_i$ ($i \in \{1, \ldots, m\}$). For a fixed target record $z = z_j$, the probability of an accurate reconstruction is

$$\Pr_{Z\sim\pi}[l(Z, t) = 0] = \pi(z) \text{ for all } z' \in \mathcal{Z}$$

In the particular case where $\pi$ is uniform, i.e., the attacker does not have any further background knowledge on the value of $t$ and the prior distribution, all reconstruction candidates $z_i$ are equally likely:

$$\Pr_{Z\sim\pi}[l(Z, t) = 0] = \frac{1}{|\mathcal{Z}|} \text{ for all } z' \in \mathcal{Z}$$

Calculating the supremum over these probabilities yields $\kappa = \max_{z\in\mathcal{Z}} \pi(z)$. Using Theorem 7.3, we arrive at the following bound:

$$\Pr_{\substack{Z\sim\pi \\ \theta\sim\mathcal{M}(D_Z)}}[l(Z, A(\theta)) \leq 0] \leq \min\{\kappa_0(\mathrm{e}^{\varepsilon}-1), \kappa_0(m-1)\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1} + \kappa_0 - \kappa_0^-\}$$

In the case of uniform priors, we get $\kappa = \frac{1}{m}$ where $m = |Z|$. Following Equation (6), we have:

$$\Pr_{\substack{Z\sim U[m] \\ \theta\sim\mathcal{M}(D_Z)}}[l(Z, A(\theta)) \leq 0] \leq \min\{\frac{1}{m}(\mathrm{e}^{\varepsilon} - 1), \frac{m-1}{m}\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\}$$

where $U[m]$ denotes the uniform distribution over $[m] = \{0, 1, \ldots, m-1\}$.

Note that an adversary that has to choose a binary attribute value has exactly the same success rate as a MIA.

The bounds obtained in this section allow us to systematize the knowledge about the attack resilience provided by DP as summarized in Table 2. We see that our results improve existing bounds. To the best of our knowledge, we are the first to prove a general bound for the attribute advantage without assumptions on the attack strategy.

## 9 Discussion and Future Work

DP mechanisms require selecting an $\varepsilon$ value that balances privacy and utility. While $\varepsilon$ can range from 0 to infinity, understanding its exact meaning can be challenging. One approach to improve interpretability is through adversarial bounds. Particularly, the bound on membership inference advantage has been extensively studied in the literature.

An advantage in an MIA ranges from -1 to 1, and its interpretation is well-understood in hypothesis testing. Translating $\varepsilon$ into an advantage bound provides a direct measure of attack protection, aiding in choosing an appropriate $\varepsilon$ value. However, other types of attacks exist, necessitating a more generalized attack performance metric and corresponding adversarial bounds.

In this context, our $(\eta, \gamma)$-U-ReRo provides a generalization of the membership and attribute advantages to arbitrary reconstruction attacks. Since we make no assumption on or restriction of data distribution or attack model, our Theorem 7.2 can be used to select the exact epsilon needed to limit the advantage of the attack according to specific requirements. $\gamma$ takes values between 0 and $1 - \mu(n, D, \pi) \leq 1$ where $\mu(n, D, \pi) = \Pr_{Z,Z_0\sim\pi}[Z = Z_0]$, i.e. the probability of resampling from the distribution $\pi$, analogous to the cases of membership and attribute advantages as defined in [35].

This allows us to directly interpret $\gamma$ as the increase in the adversary's probability of success due to the participation of a single

record. If $\gamma = 0$, there is no risk (regarding the selected attack) in participating because the probability of an adversary correctly reconstructing the record is the same as it would be without participating, e.g. random guessing. The higher the $\gamma$, the higher the risk of participation, and if $\gamma = 1 - \mu$, then the risk of participation concerning the considered attack is highest. This means that the probability of a correct reconstruction of any participant's record is 1 and the adversary always succeeds in reconstructing participant information whereas the probability of correct reconstruction is 0 for any non-participant, i.e. the reason for success in reconstruction is participation. To simplify the parameter $\gamma$ if we know $\mu$, we can normalize the value analogously to $\text{Adv}^s_{MIA}$ which is the normalized value of $\text{Adv}_{MIA}$ under uniform priors.

It follows that a randomized mechanism $\mathcal{M}$ that is $(\eta, \gamma)$-U-ReRo implies that the advantage—the difference between the adversary's probability of successful reconstruction of a participant $z$, and the probability of successful reconstruction in general without $z$'s record being part in the training—is limited by $\gamma$. The parameter $\eta$ encodes what we consider a successful reconstruction, i.e. a reconstruction $z'$ of $z$ such that the error is less than $\eta$, $l(z, z') \leq \eta$. By considering the probability of correctly reconstructing the non-participant records, we get the precise leakage generated using the information the model has learned about a participant. Therefore, by limiting $\gamma$, we control the impact of each participant in the learning mechanism concerning a selected attack.

When putting our results into practice, the first step is to determine the parameters of our use case: the distribution of the data $\pi$, the attack against which we aim to protect individuals, and what is considered a successful attack, i.e. the error function $l$ and the minimum allowed error $\eta$. For a DRA with arbitrary (non-uniform) distributions and error $\eta > 0$, we must use Theorem 7.2. For example, given that a reconstruction of 90% of an individual's DNA is considered a successful reconstruction, we set $\eta = 0.1$ and, using the distribution of the data $\pi$, we compute $\kappa_{0.1}$. Now, if the participants of the training data do not want the advantage to be greater than $\gamma = 0.2$, we can solve the formula and obtain the $\varepsilon$ value needed to train the model.

Moreover, under certain conditions, we can obtain tighter bounds. For instance, when only perfect reconstructions are considered successful, i.e., $\eta = 0$. In this case, given a non-uniform distribution, we can use Theorem 7.3 as a criterion. In addition, if the distribution is uniform, we can use the improved bound presented in Equation (6) which is directly applicable to attribute or membership inference attacks.

Note that while our theorems provide improved bounds for general and specific attacks (see Table 2), we have no general proof of the tightness of our theorems. Hence, even if the estimation we provide is better than previous work, we may still overestimate the risk, thus incurring an excessive utility loss for the privacy requirements of the participants.

In this direction, an empirical evaluation testing the gap between our theoretical bound and the empirical $\gamma$ under DP protection could be an interesting line of research to show with an empirical example that the bounds are tight, or otherwise to give an intuition about the degree of untightness of our results. This would help in the direction of a theoretical refinement of these bounds.

## 10 Conclusion

The trade-off between the utility and privacy of DP mechanisms depends on an appropriate parametrization. This choice has proven to be difficult because the effect on the actual privacy that an individual's record enjoys is not intuitive. We analyzed this problem through the lens of attack resilience, considering membership-inference (MIA), attribute-inference (AIA), and data-reconstruction (DRA) attacks.

Surveying and analyzing published attack performance metrics, we find that reconstruction robustness [1] is the only existing general metric for arbitrary attacks. We showed that the relationship between privacy parameters and this metric is not tight. This can lead to an overestimation of the privacy risk, an overestimation of the required privacy budget, and consequently to low utility results. To address this issue, we provide a tighter bound for ReRo under perfect reconstruction. We show the improvement of our bound in Figure 2.

Furthermore, we found that ReRo is not a consistent general performance metric concerning attribute and membership advantages. The cause is an overestimation of privacy leakage, explained by the privacy fallacy. We hence demonstrated this discrepancy between the values of ReRo and the advantages.

We proposed a new, more general performance metric for attacks in response: Unbiased Reconstruction Robustness (U-ReRo). This metric corrects the privacy fallacy. Our results show that, in contrast to ReRo, U-ReRo yields the established values when applied to the particular scenarios of MIAs and AIAs. Therefore, we conclude that U-ReRo is a consistent general attack performance metric, and we used it to systematize DP resilience against various attacks.

Finally, we proved several results that bound $(\eta, \gamma)$-U-ReRo under DP. The bounds directly relate the attack mitigation to the privacy parameters of DP. Therefore, we provide a criterion for selecting $\varepsilon$ for practitioners. Moreover, we highlight that our relation between U-ReRo and $\varepsilon$ allows us to select lower privacy values and thus obtain higher utility while keeping the same $\gamma$ that we would estimate with ReRo. This results directly from avoiding the risk overestimation of ReRo.

In conclusion, our new metric U-ReRo and our bounds on it under DP allow us to systematize the knowledge about attack resilience provided by DP, as summarized in Table 2. In this table, we see that our results improve on existing bounds. To the best of our knowledge, we are the first to prove a general bound for the attribute advantage without assumptions on the attack strategy. We also improve the bounds for $(0, \gamma)$-ReRo and provide a general metric U-ReRo that allows us to systematically compare the protection that DP provides against different attacks.

## Acknowledgments

This work was funded by the Topic Engineering Secure Systems of the Helmholtz Association (HGF) and supported by KASTEL Security Research Labs, Karlsruhe. It also received support from the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI).

## References


[1] Borja Balle, Giovanni Cherubin, and Jamie Hayes. 2022. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, 1138–1156. DOI: 10.1109/SP46214.2022.9833677.

[2] Daniel Bernau, Jonas Robl, Philip W Grassal, Steffen Schneider, and Florian Kerschbaum. 2021. Comparing local and central differential privacy using membership inference attacks. In *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 22–42. DOI: 10.1007/978-3-030-81242-3_2.

[3] Benjamin Bichsel, Samuel Steffen, Ilija Bogunovic, and Martin Vechev. 2021. Dp-sniper: black-box discovery of differential privacy violations using classifiers. In *2021 IEEE Symposium on Security and Privacy (SP)*, 391–409. DOI: 10.1109/SP40001.2021.00081.

[4] Mark Bun et al. 2021. Statistical inference is not a privacy violation. DifferentialPrivacy.org. https://differentialprivacy.org/inference-is-not-a-privacy-violation/. (June 2021).

[5] Chris Clifton and Tamir Tassa. 2013. On syntactic anonymity and differential privacy. In *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88–93. DOI: 10.1109/ICDEW.2013.6547433.

[6] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. 2013. Unique in the crowd: the privacy bounds of human mobility. *Scientific reports*, 3, 1–5. DOI: 10.1038/srep01376.

[7] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, and Alex "Sandy" Pentland. 2015. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science*, 347, 536–539. DOI: 10.1126/science.1256297.

[8] Damien Desfontaines and Balázs Pejó. 2020. Sok: differential privacies. *Proceedings on Privacy Enhancing Technologies*. DOI: 10.2478/popets-2020-0028.

[9] Clemens Deußer, Steffen Passmann, and Thorsten Strufe. 2020. Browsing unicity: on the limits of anonymizing web tracking data. In *Proceedings of the IEEE Symposium on Security and Privacy (S&P)*. DOI: 10.1109/SP40000.2020.00018.

[10] Zeyu Ding, Yuxin Wang, Guanhong Wang, Danfeng Zhang, and Daniel Kifer. 2018. Detecting violations of differential privacy. In *In 2018 ACM SIGSAC Conference on Computer and Communications Security (CCS '18)*, 475–489. DOI: 10.1145/3243734.3243818.

[11] Cynthia Dwork, Nitin Kohli, and Deirdre Mulligan. 2019. Differential privacy in practice: expose your epsilons! *Journal of Privacy and Confidentiality*, 9. DOI: 10.29012/jpc.689.

[12] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284. DOI: 10.1007/11681878_14.

[13] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9, 3–4, 211–407. DOI: 10.1561/0400000042.

[14] Úlfar Erlingsson, Ilya Mironov, Ananth Raghunathan, and Shuang Song. 2019. That which we call private. *arXiv preprint arXiv:1908.03566*. DOI: 10.48550/arXiv.1908.03566.

[15] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1322–1333. DOI: 10.1145/2810103.2813677.

[16] Trung Ha, Trang Vo, Tran Khanh Dang, and Nguyen Thi Huyen Trang. 2023. Differential privacy under membership inference attacks. In *International Conference on Future Data and Security Engineering*, 255–269. DOI: 10.1007/978-981-99-8296-7_18.

[17] Andreas Haeberlen, Benjamin C. Pierce, and Arjun Narayan. 2011. Differential privacy under fire. In *Proceedings of the 20th USENIX Conference on Security*, 33. DOI: 10.5555/2028067.2028100.

[18] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. 2023. Investigating membership inference attacks under data dependencies. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, 473–488. DOI: 10.1109/CSF57540.2023.00013.

[19] Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical disclosure control*. John Wiley & Sons. DOI: 10.1002/9781118348239.

[20] Bargav Jayaraman and David Evans. 2019. Evaluating differentially private machine learning in practice. In *28th USENIX Security Symposium (USENIX Security 19)*, 1895–1912.

[21] Daniel Kifer, John M. Abowd, Robert Ashmead, Ryan Cumings-Menon, Philip Leclerc, Ashwin Machanavajjhala, William Sexton, and Pavel Zhuravlev. 2022. Bayesian and Frequentist Semantics for Common Variations of Differential Privacy: Applications to the 2020 Census. Tech. rep. eprint arXiv. eprint: 2209.03310.

[22] Daniel Kifer and Ashwin Machanavajjhala. 2011. No free lunch in data privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*, 193–204. DOI: 10.1145/1989323.1989345.

[23] Jaewoo Lee and Chris Clifton. 2011. How much is enough? choosing $\varepsilon$ for differential privacy. In *Information Security: 14th International Conference, ISC 2011, Xi'an, China, October 26-29, 2011. Proceedings 14*, 325–340. DOI: 10.1007/978-3-642-24861-0_22.

[24] David JC MacKay. 2003. *Information theory, inference and learning algorithms*. Cambridge university press.

[25] Priyanka Nanayakkara, Mary Anne Smart, Rachel Cummings, Gabriel Kaptchuk, and Elissa M Redmiles. 2023. What are the chances? explaining the epsilon parameter in differential privacy. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1613–1630.

[26] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust de-anonymization of large sparse datasets. In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, 111–125. DOI: 10.1109/SP.2008.33.

[27] Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. 2021. Adversary instantiation: lower bounds for differentially private machine learning. In *2021 IEEE Symposium on security and privacy (SP)*, 866–882. DOI: 10.1109/SP40001.2021.00069.

[28] Natalia Ponomareva, Hussein Hazimeh, Alex Kurakin, Zheng Xu, Carson Denison, H Brendan McMahan, Sergei Vassilvitskii, Steve Chien, and Abhradeep Guha Thakurta. 2023. How to dp-fy ml: a practical guide to machine learning with differential privacy. *Journal of Artificial Intelligence Research*, 77, 1113–1201. DOI: 10.1613/jair.1.14649.

[29] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, 3–18. DOI: 10.48550/arXiv.1610.05820.

[30] Latanya Sweeney. 2000. Simple demographics often identify people uniquely. *Carnegie Mellon University, Data Privacy*.

[31] Julian Todt, Simon Hanisch, and Thorsten Strufe. 2024. Fantômas: understanding face anonymization reversibility. *Proceedings on Privacy Enhancing Technologies*, 2024, 4, 24–43. DOI: 10.56553/popets-2024-0105.

[32] Jincheng Wang, Zhuohua Li, John C.S. Lui, and Mingshen Sun. 2022. Topology-theoretic approach to address attribute linkage attacks in differential privacy. *Computers & Security*, 113, 102552. DOI: 10.1016/j.cose.2021.102552.

[33] Xi Wu, Matthew Fredrikson, Somesh Jha, and Jeffrey F Naughton. 2016. A methodology for formalizing model-inversion attacks. In *2016 IEEE 29th computer security foundations symposium (CSF)*, 355–370. DOI: 10.1109/CSF.2016.32.

[34] Dayong Ye, Sheng Shen, Tianqing Zhu, Bo Liu, and Wanlei Zhou. 2022. One parameter defense-defending against data inference attacks via differential privacy. *IEEE Transactions on Information Forensics and Security*, 17, 1466–1480. DOI: 10.1109/TIFS.2022.3163591.

[35] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. 2018. Privacy risk in machine learning: analyzing the connection to overfitting. In *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282. DOI: 10.1109/CSF.2018.00027.

[36] Oualid Zari, Javier Parra-Arnau, Ayşe Ünsal, Thorsten Strufe, and Melek Önen. 2022. Membership inference attack against principal component analysis. In *International Conference on Privacy in Statistical Databases*, 269–282. DOI: 10.1007/978-3-031-13945-1_19.

[37] Qiuchen Zhang, Jing Ma, Yonghui Xiao, Jian Lou, and Li Xiong. 2020. Broadening differential privacy for deep learning against model inversion attacks. In *2020 IEEE International Conference on Big Data (Big Data)*, 1061–1070. DOI: 10.1109/BigData50022.2020.9378274.

# A Appendix of Proofs

In this section, we provide all the mathematical proofs of the results presented in this paper. We also prove some auxiliary lemmas that allow us to complete the main proofs.

All the proofs use the notation of discrete random variables. To obtain the analogous results for the continuous case, one must replace $\Pr(\mathcal{M}(D) = O)$ by the density or probability function and the sums by integrals.

Lemma A.1. *Given $\varepsilon, A, B \in \mathbb{R}_{\geq 0}$. If the following inequalities are satisfied:*

$$(a)\, A \leq \mathrm{e}^{\varepsilon} B \quad \textit{and} \quad (b)\, (1-B) \leq \mathrm{e}^{\varepsilon}(1-A)$$

*then, $A - B \leq \frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}$.*

Proof. Using (a) we obtain:

$$A \leq \mathrm{e}^{\varepsilon} B \Leftrightarrow 1 - A \geq 1 - \mathrm{e}^{\varepsilon} B \Leftrightarrow (1-A) + \mathrm{e}^{\varepsilon} B \geq 1$$

Analogously, using (b) we obtain:

$$(1-B) \leq \mathrm{e}^{\varepsilon}(1-A) \Leftrightarrow 1 - 1 + B \geq 1 - \mathrm{e}^{\varepsilon}(1-A)$$
$$\Leftrightarrow B + \mathrm{e}^{\varepsilon}(1-A) \geq 1$$

Summing both equations we obtain:

$$(1-A) + \mathrm{e}^{\varepsilon} B + B + \mathrm{e}^{\varepsilon}(1-A) \geq 2 \Leftrightarrow$$
$$(1+\mathrm{e}^{\varepsilon})(1 - A + B) \geq 2 \Leftrightarrow$$
$$A - B \leq \frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}$$

□

Lemma A.2. *Given $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ an $\varepsilon$-DP learning mechanism. For any error function $l\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ and for any pair of records $z_1, z_2 \in \mathcal{Z}$, we obtain:*

$$\Pr_{\theta \sim \mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) \leq \eta] \leq \mathrm{e}^{\varepsilon} \Pr_{\theta \sim \mathcal{M}(D_{z_2})}[l(z_1, A(\theta)) \leq \eta]$$

*and,*

$$1 - \Pr_{\theta \sim \mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) \leq \eta] \leq \mathrm{e}^{\varepsilon}\left(1 - \Pr_{\theta \sim \mathcal{M}(D_{z_2})}[l(z_1, A(\theta)) \leq \eta]\right)$$

Proof. The first inequality follows directly from the definition of $\varepsilon$-DP:

$$\Pr_{\theta \sim \mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) \leq \eta] = \sum_{O \in \Theta} \mathbf{1}_{\{l(z_1, A(O)) \leq \eta\}} \Pr(\mathcal{M}(D_{z_1}) = O)$$
$$\leq \sum_{O \in \Theta} \mathbf{1}_{\{l(z_1, A(O)) \leq \eta\}} \mathrm{e}^{\varepsilon} \Pr(\mathcal{M}(D_{z_2}) = O)$$
$$= \mathrm{e}^{\varepsilon} \sum_{O \in \Theta} \mathbf{1}_{\{l(z_1, A(O)) \leq \eta\}} \Pr(\mathcal{M}(D_{z_2}) = O)$$
$$= \mathrm{e}^{\varepsilon} \Pr_{\theta \sim \mathcal{M}(D_{z_2})}[l(z_1, A(\theta)) \leq \eta]$$

Where $\mathbf{1}_{\{l(z_1, A(O)) \leq \eta\}}$ the characteristic function that outputs 1 when the condition $l(z_1, A(O)) \leq \eta$ is satisfied and 0 otherwise. The second inequality is derived from

$$1 - \Pr_{\theta \sim \mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) \leq \eta] = \Pr_{\theta \sim \mathcal{M}(D_{z_1})}[l(z_1, A(\theta)) > \eta],$$

applying the $\varepsilon$-DP condition analogously to the first inequality. □

Proposition 5.1 (Improved Bound for ReRo against perfect reconstruction). *Let $\pi$ be a prior over $\mathcal{Z}$, $l : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function, $|\mathcal{Z}| = m$ and $\kappa_0 = \kappa_{\pi,l}(0)$. If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ satisfies $\varepsilon$-DP, then it also satisfies $(0, \gamma)$-ReRo with $\gamma = \min\{\kappa_0 \mathrm{e}^{\varepsilon}, \kappa_0\left(1 + (m-1)\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\right)\}$, i.e.,*

$$\Pr_{\substack{z \sim \pi \\ \theta \sim \mathcal{M}(D_z)}}[l(z, A(\theta)) = 0] \leq \min\{\kappa_0 \mathrm{e}^{\varepsilon}, \kappa_0\left(1 + (m-1)\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\right)\}$$

Proof. First, we apply the definition of $\kappa_0 \equiv \kappa_{\pi,l}(0)$ to the case of perfect reconstruction,

$$\kappa_0 \coloneqq \sup_{z \in \mathcal{Z}} \Pr_{Z' \sim \pi}(l(Z', z) = 0) = \sup_{z \in \mathcal{Z}} \Pr_{Z' \sim \pi}(Z' = z) = \sup_{z \in \mathcal{Z}} \pi(z),$$

where $\pi(z)$ is the probability of drawing $z$ from $\pi$.

Then, applying the $\varepsilon$-DP condition to the definition of ReRo, we obtain:

$$\Pr_{\substack{Z \sim \pi \\ \theta \sim \mathcal{M}(D_Z)}}(l(A(\theta), Z) = 0) = \sum_{z \in \mathcal{Z}} \Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0)\pi(z)$$
$$= \sum_{z \neq z_0} \Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0)\pi(z) + \Pr_{\theta \sim \mathcal{M}(D_{z_0})}(l(A(\theta), z_0) = 0)\pi(z_0)$$
$$\leq \sup_{z \in \mathcal{Z}} \pi(z)\left(\sum_{z \neq z_0} \Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0) + \Pr_{\theta \sim \mathcal{M}(D_{z_0})}(l(A(\theta), z_0) = 0)\right)$$
$$= \kappa_0\left(\sum_{z \neq z_0} \Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0) + 1 - \Pr_{\theta \sim \mathcal{M}(D_{z_0})}(l(A(\theta), z_0) > 0)\right)$$
$$= \kappa_0\left(\sum_{z \neq z_0} \Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0) + 1 - \sum_{z \neq z_0} \Pr_{\theta \sim \mathcal{M}(D_{z_0})}(l(A(\theta), z) = 0)\right)$$
$$= \kappa_0\left(1 + \sum_{z \neq z_0}\left[\Pr_{\theta \sim \mathcal{M}(D_z)}(l(A(\theta), z) = 0) - \Pr_{\theta \sim \mathcal{M}(D_{z_0})}(l(A(\theta), z) = 0)\right]\right)$$
$$\overset{(*)}{\leq} \kappa_0\left(1 + \sum_{z \neq z_0} \frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\right) = \kappa_0\left(1 + (m-1)\frac{\mathrm{e}^{\varepsilon}-1}{\mathrm{e}^{\varepsilon}+1}\right)$$

where $(*)$ follows from direct application of Lemma A.1 and Lemma A.2. □

Proposition 6.1. *Given a learning mechanism $\mathcal{M}$, consider a strong membership experiment (Def. 4.8) where $D_-$ is known and the two possible members $\{z_0, z_1\}$ are uniformly distributed. Given $l$, the membership error function (Eq. 7), such that $l(z, z') = 0$ if $z = z'$ and 1 otherwise, we obtain:*

$$\mathrm{Adv}^{s}_{\mathrm{MIA}} = 2 \Pr_{\substack{z_i \sim U\{0,1\} \\ \theta \sim \mathcal{M}(D_{z_i})}}[l(z_i, A(\theta)) = 0] - 1.$$

Proof. First, we can rewrite the strong advantage of a MIA for the strong MIA experiment (Def. 4.8) as follows:

$$\mathrm{Adv}^{s}_{\mathrm{MIA}} \coloneqq 2\Pr[\mathrm{Exp}^{\mathrm{MIA}}_{s}] - 1$$
$$= 2\left(\frac{1}{2} \Pr_{\theta \sim \mathcal{M}(D_0)}(A(\theta) = z_0) + \frac{1}{2} \Pr_{\theta \sim \mathcal{M}(D_1)}(A(\theta) = z_1)\right) - 1$$

On the other hand since $\pi = U\{0, 1\}$, and $\mathcal{Z} = \{0, 1\}$ we have that,

$$\Pr_{\substack{z \sim U\{0,1\} \\ \theta \sim \mathcal{M}(D_z)}}[l(z, A(\theta)) \leq 0] = \sum_{z=0}^{1} \Pr_{\theta \sim \mathcal{M}(D_z)}[l(z, A(\theta)) = 0] \cdot \frac{1}{2}$$
$$= \frac{1}{2}\left(\Pr_{\theta \sim \mathcal{M}(D_0\})}(A(\theta) = z_0) + \Pr_{\theta \sim \mathcal{M}(D_1\})}(A(\theta) = z_1)\right)$$

Therefore, we have that:

$$\mathrm{Adv}^{s}_{\mathrm{MIA}} = 2 \Pr_{\substack{z\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_z)}}[l(z,A(\theta))=0] - 1 \tag{8}$$

$$\Leftrightarrow \Pr_{\substack{z\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_z)}}[l(z,A(\theta))\le 0] = \frac{1}{2}(\mathrm{Adv}_{\mathrm{MIA}}+1)$$

□

Theorem 7.2 ($\varepsilon$-DP Implies $(\eta,\gamma)$-U-ReRo). *Let $\pi, l$ and $\eta \ge 0$ follow Definition 7.1, and $\kappa_\eta \equiv \kappa_{\pi,l}(\eta)$. If a mechanism $\mathcal{M}$ satisfies $\varepsilon$-DP, then it also satisfies $(\eta,\gamma)$-U-ReRo with*

$$\gamma = \min\{\kappa_\eta(\mathrm{e}^\varepsilon - 1), \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1}\}$$

Proof. Following Lemma A.1 we denote:

- $A \equiv \Pr_{\theta\sim\mathcal{M}(D_z)}^{z\sim\pi}[l(z,A(\theta))\le\eta]$
- $B \equiv \mathbb{E}_{Z_0\sim\pi}\left[\Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi} l(Z,A(\theta))\le\eta\right]$.

Using this notation, by definition of $(\eta,\gamma)$-U-ReRo, we need to prove that $\gamma = A - B \le \min\{\kappa_\eta(\mathrm{e}^\varepsilon-1), \frac{\mathrm{e}^\varepsilon-1}{\mathrm{e}^\varepsilon+1}\}$.

Applying Lemma A.2 to $B$ we obtain:

$$\begin{aligned} B &\equiv \mathbb{E}_{Z_0\sim\pi} \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta] \\ &= \sum_{z_0\in\mathcal{Z}} \Pr_{\theta\sim\mathcal{M}(D_{z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta]\,\pi(z_0) \\ &\overset{(A.2)}{\ge} \mathrm{e}^{-\varepsilon}\sum_{z_0\in\mathcal{Z}} \Pr_{\theta\sim\mathcal{M}(D_Z)}^{Z\sim\pi}[l(Z,A(\theta))\le\eta]\,\pi(z_0) \\ &= \mathrm{e}^{-\varepsilon}A\sum_{z_0\in\mathcal{Z}}\pi(z_0) = \mathrm{e}^{-\varepsilon}A. \end{aligned}$$

In addittion,, Proposition 4.16 [1] states:

$$A \equiv \Pr_{\theta\sim\mathcal{M}(D_z)}^{z\sim\pi}[l(z,A(\theta))\le\eta] \le \kappa_{\pi,l}(\eta)\mathrm{e}^\varepsilon \equiv \kappa_\eta\mathrm{e}^\varepsilon$$

Therefore, aggregating both results we obtain:

$$A - B \le A - \mathrm{e}^{-\varepsilon}A = A(1-\mathrm{e}^{-\varepsilon}) \overset{(4.16)}{\le} \kappa_{\pi,l}(\eta)\mathrm{e}^\varepsilon(1-\mathrm{e}^{-\varepsilon}). \tag{9}$$

Applying Lemma A.2 to $A$, we obtain:

$$\begin{aligned} A &\equiv \Pr_{\theta\sim\mathcal{M}(D_z)}^{z\sim\pi}[l(z,A(\theta))\le\eta] \\ &= \sum_{z\in\mathcal{Z}} \Pr_{\theta\sim\mathcal{M}(D_z)}[l(z,A(\theta))\le\eta]\pi(z) \\ &\overset{(A.2)}{\le} \sum_{z\in\mathcal{Z}} \mathrm{e}^\varepsilon \Pr_{\theta\sim\mathcal{M}(D_{z_0})}[l(z,A(\theta))\le\eta]\pi(z) \\ &= \mathrm{e}^\varepsilon \Pr_{\theta\sim\mathcal{M}(D_{z_0})}^{z\sim\pi}[l(z,A(\theta))\le\eta] \end{aligned}$$

Since this is true for all $z_0 \in \mathcal{Z}$, applying the properties of the expected value directly, we derive:

$$A \le \mathbb{E}\left[\mathrm{e}^\varepsilon \Pr_{\theta\sim\mathcal{M}(D_{z_0})}^{z\sim\pi}[l(z,A(\theta))\le\eta]\right] = \mathrm{e}^\varepsilon B.$$

Using Lemma A.2 again, we obtain that for all $z_0 \in \mathcal{Z}$,

$$\begin{aligned} 1 - \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta] &= \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))>\eta] \\ &\overset{(A.2)}{\le} \mathrm{e}^\varepsilon \Pr_{\theta\sim\mathcal{M}(D_Z)}^{Z\sim\pi}[l(Z,A(\theta))>\eta] \\ &= \mathrm{e}^\varepsilon\left(1 - \Pr_{\theta\sim\mathcal{M}(D_Z)}^{Z\sim\pi}[l(Z,A(\theta))\le\eta]\right) \end{aligned}$$

therefore,

$$\begin{aligned} 1 - B &= 1 - \mathbb{E}_{Z_0\sim\pi} \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta] \\ &= \mathbb{E}_{Z_0\sim\pi}(1) - \mathbb{E}_{Z_0\sim\pi} \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta] \\ &= \mathbb{E}_{Z_0\sim\pi}\left(1 - \Pr_{\theta\sim\mathcal{M}(D_{Z_0})}^{Z\sim\pi}[l(Z,A(\theta))\le\eta]\right) \\ &\le \mathrm{e}^\varepsilon\left(1 - \Pr_{\theta\sim\mathcal{M}(D_Z)}^{Z\sim\pi}[l(Z,A(\theta))\le\eta]\right) = \mathrm{e}^\varepsilon(1-A) \end{aligned}$$

So, applying Lemma A.1, we have that

$$A - B \le \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + 1}. \tag{10}$$

Aggregating the two bounds of Equation (9) and Equation (10), we obtain the result.

□

Theorem 7.3 ($\varepsilon$-DP Implies $(0,\gamma)$-U-ReRo). *Let $\pi$ and $l$ follow Definition 7.1, $|\mathcal{Z}| = m$ and $\kappa_0 = \kappa_{\pi,l}(0)$. If a mechanism $\mathcal{M}$ satisfies $\varepsilon$-DP, then it also satisfies $(0,\gamma)$-ReRo with*

$$\gamma = \min\{\kappa_0(\mathrm{e}^\varepsilon - 1), \kappa_0(m-1)\frac{\mathrm{e}^\varepsilon-1}{\mathrm{e}^\varepsilon+1} + \kappa_0 - \kappa_0^-\},$$

*where, $\kappa_0^- \equiv \kappa^-_{\pi,l}(0) \coloneqq \inf_{z'\in\mathcal{Z}}\Pr_{z\sim\pi}[l(z,z')=0]$ is the lower baseline error.*

Proof. $\gamma \le k_0(\mathrm{e}^\varepsilon - 1)$ follows directly from previous Theorem 7.2 applied to $\eta = 0$. Therefore, we just need to prove that $\gamma \le k_0(m-1)\frac{\mathrm{e}^\varepsilon-1}{\mathrm{e}^\varepsilon+1} + k_0 - k_0^-$ and we obtain the result.

From Proposition 5.1 we know that:

$$\Pr_{z\sim\pi,\theta\sim\mathcal{M}(D_-\cup\{z\})}[l(z,A(\theta))\le\eta] \le \kappa_0\left(\frac{\mathrm{e}^\varepsilon-1}{\mathrm{e}^\varepsilon+1}+1\right)$$

Since $A(\theta)\in\mathcal{Z}$, by definition of $\kappa_0^-$ we have that:

$$\mathbb{E}_{z_0\sim\pi}\left[\Pr_{\theta\sim\mathcal{M}(D_{z_0})}^{z\sim\pi} l(z,A(\theta))=0\right] \ge \mathbb{E}_{z_0\sim\pi}\left[\kappa_0^-\right] = \kappa_0^-.$$

Joining both inequalities we obtain the result. □

Proposition 8.1 ($\text{Adv}_{\text{MIA}} \Leftrightarrow$ U-ReRo). *Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a mechanism. For $A$ an informed MIA, it holds that*

$$\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \Longleftrightarrow \text{Adv}_{\text{MIA}}(A,\mathcal{M},\pi^n) \leq \gamma,$$

*Additionally, if $A^s$ is a strong MIA under uniform priors (Def. 4.7), then*

$$\mathcal{M} \text{ is } (0,\frac{\gamma}{2})\text{-U-ReRo} \Longleftrightarrow \text{Adv}^s_{\text{MIA}}(A,\mathcal{M},\pi^n) \leq \gamma.$$

Proof. Following Section 4.1

$$\text{Adv}_{MIA} = \underbrace{\Pr_{\substack{D\sim\pi^n\\ \theta\sim\mathcal{M}(D)}} \Pr_{Z\sim D}(A(\theta,Z,\pi)=0)}_{A} - \underbrace{\Pr_{\substack{D\sim\pi^n\\ \theta\sim\mathcal{M}(D)}} \Pr_{Z\sim\pi}(A(\theta,Z,\pi)=0)}_{B}$$

Now, following Definition 4.4, since our adversary is informed and we assume $D_-$ to be known, we model $\pi^n$ such that $\pi^n(D) = \pi(z)$ for all $D = D_- \cup \{z\} \equiv D_z$ , $i \in \{0,1\}$ and 0 otherwise. When $b = 0$, we challenge $A$ with the real missing training record $z$, and when $b = 1$, we challenge $A$ with a random point of the distribution $z \sim \pi$. Therefore,

$$\begin{aligned} A &= \Pr_{\substack{D\sim\pi^n\\ \theta\sim\mathcal{M}(D)}} \Pr_{Z\sim D}(A(\theta,Z,\pi)=0) \\ &= \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_z)}}(A(\theta,z,\pi)=0) \\ = \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_z)}}(A(\theta)=z) &= \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_z)}}(l(A(\theta),Z)=0) \end{aligned}$$

Therewith, we have that,

$$\begin{aligned} B &= \Pr_{\substack{D\sim\pi^n\\ \theta\sim\mathcal{M}(D)}} \Pr_{Z\sim\pi}(A(\theta,Z,\pi)=0) \\ = \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}} \Pr_{Z'\sim\pi}(A(\theta,Z',\pi)=0) &= \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}} \Pr_{Z'\sim\pi}(A(\theta)=Z') \\ = \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D)}} \Pr_{Z'\sim\pi}(l(A(\theta),Z')=0) = \mathbb{E}_{Z\in\pi} &\Pr_{\substack{Z'\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}}(l(A(\theta),Z')=0), \end{aligned}$$

which lead us to the desired result.

For the case of a strong membership experiment, following Definition 4.7, with $\pi = \text{U}\{0,1\}$ we obtain,

$$\mathbb{E}_{Z'\sim\pi}\left(\Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_{Z'})}}(l(A(\theta),Z)=0)\right) = \tag{11}$$

$$= \mathbb{E}_{Z'\sim\pi}\left(\frac{1}{2}\Pr_{\theta\sim\mathcal{M}(D_{Z'})}(A(\theta)=z_0) + \frac{1}{2}\Pr_{\theta\sim\mathcal{M}(D_{Z'})}(A(\theta)=z_1)\right) = \frac{2}{4} = \frac{1}{2} \tag{12}$$

Additionally, Proposition 8.1 states that,

$$\Pr_{\substack{z\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_z)}}[l(z,A(\theta))\leq 0] = \frac{1}{2}(\text{Adv}^s_{\text{MIA}}+1)$$

Combining both results we have that,

$$\Pr_{\substack{z\sim\pi\\ \theta\sim\mathcal{M}(D_z)}}[l(z,A(\theta))\leq\eta] - \mathbb{E}_{z_0\sim\pi}\left[\Pr_{\substack{z\sim\pi\\ \theta\sim\mathcal{M}(D_{z_0})}} l(z,A(\theta))\leq\eta\right] = \tag{13}$$

$$\frac{1}{2}(\text{Adv}^s_{\text{MIA}}+1) - \frac{1}{2} = \frac{1}{2}\text{Adv}^s_{\text{MIA}} \overset{(Appendix\ B)}{=} \text{Adv}_{\text{MIA}} \tag{14}$$

□

Proposition 8.2 ($\text{Adv}_{\text{AIA}} \Leftrightarrow$ U-ReRo). *Let $\mathcal{M}\colon \mathcal{Z}^n \to \Theta$ be a mechanism. For all data distributions $\pi^n$ and for all informed AIA attacks $A$ that know $D_-$ and try to guess the attribute of a target record $z$ from which $\phi(z)$ is known, we have*

$$\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \Longleftrightarrow \text{Adv}_{\text{AIA}}(A,\mathcal{M},\pi^n) \leq \gamma \text{ for all } A.$$

Proof. Note that, an informed AIA experiment consists of the following steps:

(1) Sample $z' \sim \pi$.
(2) Sample $b$ uniformly from $\{0,1\}$.
(3) If $b = 0$ draw $z = z'$, else $z \sim \pi$.
(4) Run $A(\phi(z),\mathcal{M}(D_{z'}),\pi) \equiv A(\phi(z),\mathcal{M}(D_{z'}))$.

Therefore,

$$\begin{aligned} \Pr(Exp^{\mathcal{A}}=1|b=0) &= \sum_{z'\in\mathcal{Z}} \Pr(A(\phi(z'),\theta)=\varphi(z')\pi(z') \\ &= \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}}(l(A(\phi(Z),\theta),\varphi(Z))=0) \end{aligned}$$

and,

$$\begin{aligned} \Pr(Exp^{\mathcal{A}}=1|b=1) &= \sum_{z'\in\mathcal{Z}} \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_Z)}}(A(\phi(z'),\theta)=\varphi(z')\pi(z') \\ &= \sum_{z'\in\mathcal{Z}} \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_{z'})}}(l(A(\phi(Z),\theta),\varphi(Z))=0)\pi(z') \\ &= \mathbb{E}_{Z'\sim\pi} \Pr_{\substack{Z\sim\pi\\ \theta\sim\mathcal{M}(D_{z'})}}(l(A(\phi(Z),\theta),\varphi(Z))=0) \end{aligned}$$

Therefore,

$$\begin{aligned} &\mathcal{M} \text{ is } (0,\gamma)\text{-U-ReRo} \Leftrightarrow \\ &\Leftrightarrow \Pr(Exp^{\mathcal{A}}=1|b=0) - \Pr(Exp^{\mathcal{A}}=1|b=1) \leq \gamma \\ &\Leftrightarrow \text{Adv}_{\text{AIA}} \leq \gamma \end{aligned}$$

□

## B Relation between Membership Advantages

The membership advantage presented in [35] takes the prior distribution into account and measures the leakage that the mechanism trained on $D$ produces by removing the effect of the prior distribution knowledge on the attack. When the prior is uniform, it does not provide any advantage beyond the real privacy leakage, and the advantage can be simplified.

Since we assume $D_-$ to be known, we model $\pi^n$ such that $\pi^n(D) = \pi(z_i) = \frac{1}{2}$ for all $D = D_- \cup \{z_i\} \equiv D_{z_i}$ , $i \in \{0,1\}$ and 0 otherwise. Therefore;

$$\begin{aligned} \text{Adv}_{MIA} &= \Pr_{\substack{D\sim\pi^n\\ \theta\sim\mathcal{M}(D)}}\left(\Pr_{Z\sim D}(A(\theta,Z,\pi)=0) - \Pr_{Z\sim\pi}(A(\theta,Z,\pi)=0)\right) \\ &= \Pr_{\substack{z_i\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_i})}} \Pr(A(\theta,z_i,\pi)=0) - \Pr_{\substack{z_i\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_i})}} \Pr_{Z\sim U\{0,1\}}(A(\theta,Z,\pi)=0) \end{aligned}$$

Where,

$$\begin{aligned} &\Pr_{\substack{z_i\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_i})}}(\Pr(A(\theta,z_i,\pi)=0) \\ &= \frac{1}{2}\Pr_{\theta\sim\mathcal{M}(D_{z_0})}(A(z_0,\theta,\pi)=0) + \frac{1}{2}\Pr_{\theta\sim\mathcal{M}(D_{z_1})}(A(z_1,\theta,\pi)=0)) \end{aligned}$$

and,

$$\Pr_{\substack{z_i\sim U\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_i})}}\Pr_{Z\sim U\{0,1\}}(A(\theta,Z,\pi)=0)$$
$$=\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_0})}(A(z_0,\theta,\pi)=0)+\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_1})}(A(z_1,\theta,\pi)=0))+$$
$$\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_0})}(A(z_1,\theta,\pi)=0)+\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_1})}(A(z_0,\theta,\pi)=0))$$

Therefore,

$$\mathrm{Adv}_{\mathrm{MIA}}=$$
$$=\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_0})}(A(z_0,\theta,\pi)=0)+\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_1})}(A(z_1,\theta,\pi)=0))-$$
$$\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_0})}(A(z_1,\theta,\pi)=0)-\frac{1}{4}\Pr_{\theta\sim\mathcal{M}(D_{z_1})}(A(z_0,\theta,\pi)=0))$$
$$=\frac{1}{2}\Pr_{\substack{\alpha\in\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_\alpha})}}(A^s(\theta)=\alpha)-\frac{1}{2}\Pr_{\substack{i\in\{0,1\}\\ \theta\sim\mathcal{M}(D_{z_\alpha})}}(A^s(\theta)\neq\alpha)$$
$$=\frac{1}{2}\mathrm{Adv}^s_{\mathrm{MIA}}$$