UNDERSTANDING DISCLOSURE RISK IN DIFFERENTIAL PRIVACY WITH APPLICATIONS TO NOISE CALIBRATION AND AUDITING

Extended version

PATRICIA GUERRA-BALBOA, ANNIKA SAUER, HÉBER H. ARCOLEZI, AND THORSTEN STRUFE

Abstract. Differential Privacy (DP) has emerged as the leading framework for protecting sensitive data in analysis, yet a key challenge lies in understanding its practical implications, particularly in terms of attack resilience. While significant progress has been made in studying membership inference attacks (MIAs), they capture only a threat of the broader adversarial landscape. Recently, data reconstruction attacks (DRAs) have been proposed as a unifying framework, providing a systematic way to formalize and analyze diverse attacker models under a common metric: Reconstruction robustness (ReRo).

In this work, we provide empirical evidence of ReRo limitations, leading to misleading risk estimates and violations of the existing bounds when tested in real cases. To overcome these limitations we propose *reconstruction advantage*, a new consistent metric across MIAs, AIAs, and DRAs, providing a unified framework understanding DP attack resilience. Moreover, we derive tight bounds formalizing a precise connection between DP noise and adversarial advantage, enabling principled privacy-budget calibration. Moreover, we provide the optimal attack strategy against any DP mechanism, attacker goal and auxiliary knowledge. We show the impact of our new framework in both DP noise calibration–improving utility– and DP auditing–broadening the scope and improving accuracy over existing LDP auditing tools.

1. Introduction

Differential privacy (DP) [13] and its distributed variant, local DP (LDP), have emerged as the de facto standard to mitigate privacy risk, i.e., the extent to which a learning process allows sensitive information about participants to be inferred. Hence, DP aims to make participation as safe as not participating [12]. The privacy-utility trade-off in DP is governed by the privacy budget ε, which bounds the privacy loss – smaller values providing stronger guarantees – and by δ, which captures the probability mass of outcomes in which the guarantee may fail, weighted by the severity of their deviation from ε [41]. Despite its solid theoretical foundation, a central practical question remains: How do these formal parameters – especially ε – translate into concrete protection against real-world attacks? [44] This question is critical for calibrating ε: if set too high, sensitive information may be exposed; if too low, utility is unnecessarily compromised. Moreover, the relation between DP and attack resilience is crucial for DP auditing, i.e., methods to estimate empirical privacy [25], test tightness of DP mechanisms [46], and to detect bugs [54].

In this regard, major progress has been made in connecting DP to the risk of *membership inference attacks* (MIAs) [15, 24, 59], even bypassing the step of choosing ε trough direct noise calibration for desired MIA risk level [32]. However, MIA represents only one aspect of the broader privacy risk landscape. In particular, *attribute inference attacks* (AIA) [59]–which can target sensitive information even when membership is public [5]–are less understood, and current analyses provide limited insight into how DP mitigates them. Recently, *data reconstruction attacks* (DRAs) [5] were presented as a unifying framework that subsumes AIAs and MIAs, while also accounting for the impact of partial or imperfect reconstruction as privacy breach–e.g., exposing a car's license plate in an image sufficiently harms privacy even if the background is inaccurate.

Inria Centre at the University Grenoble Alpes
Karlsruhe Institute of Technology, KASTEL SRL
E-mail addresses: patricia.balboa@kit.edu, annika.sauer@student.kit.edu, heber.hwang-arcolezi@inria.fr, thorsten.strufe@kit.edu.

2020 *Mathematics Subject Classification.* 68P27.

Balle, Cherubin, and Hayes [5] introduced the first metric for DRAs, *reconstruction robustness* (ReRo), a pioneering work that, for the first time, provided a unified view of DP attack resilience by aggregating various risks in a single framework. While this contribution was foundational and highly relevant as an initial step toward unifying attack resilience, ReRo suffers from critical limitations that undermine its reliability as a comprehensive adversarial metric.

First, ReRo and all existing bounds [5, 22] only apply to attackers without target-specific auxiliary knowledge, i.e., it does not consider any partial information about the record, such as demographic attributes or social media information, to be public, excluding any AIA [17, 59]. This restriction is highly unrealistic, since most real-world privacy attacks historically rely on publicly available information about the victim [42, 45, 52]. We empirically confirm this limitation: when target-specific auxiliary information is included, existing ReRo bounds fail (see Figure 4). Second, ReRo, is a success probability. As a result, ReRo–and any bounds derived from it such as [5, 20, 22]–penalizes a mechanism for providing global statistical knowledge, which is precisely the purpose of data release, and confuses participation risk with success arising from background knowledge or statistical imputation, as we show in Table 1 and Figure 4. Crucially, such success would occur even if the learning process had not taken place and hence do not constitute a genuine risk [8, 31]. Since parameter choices directly determine the privacy–utility trade-off, this misinterpretation leads to unnecessary utility loss, as we illustrate in Figure 2.

As a solution, we introduce *η-reconstruction advantage* (η-RAD), which extends existing advantage metrics to the unifying DRA framework and overcomes the mentioned limitations of ReRo. Moreover, we establish tight bounds linking the noise of DP mechanisms to η-RAD. This connection allows us to understand the user's participation risk: given a specific mechanism, one can quantify the exact risk of private information disclosure that a participant assumes. First, in Theorem 4.2, we establish a closed-form bound that holds independently of the adversary's auxiliary knowledge, thereby serving as a worst-case guarantee when the nature of the target-specific auxiliary knowledge is unknown. Second, in Theorem 4.3 we prove an auxiliary-dependent bound that allows calibration directly from the mechanism, bypassing the privacy parameters, and taking the auxiliary knowledge into account.

To assess the tightness of these bounds, we construct and prove the optimal attack strategy for any attack goal, auxiliary knowledge, and DP mechanism–a result that, to the best of our knowledge, is unprecedented and can be further used to empirically test risk in future analyses. We then instantiate this optimal attack strategy on real datasets and mechanisms, demonstrating the tightness of both our bounds in practice, but in different senses–Theorem 4.3 is universally tight – given any mechanism and auxiliary knowledge, there exists an attack that attains the bound, while Theorem 4.2 tightness holds when the attacker auxiliary knowledge is unknown.

While Theorem 4.3 is universally tight, hence it can not be further improved, it is a white-box bound, i.e., implies complete knowledge of the mechanism $\mathcal{M}$. This limits its applicability to black-box DP auditing [37]. Hence, we provide a closed-form black-box upper bounds under additional assumptions, such as the reconstruction setting without auxiliary knowledge—i.e., when the whole target record is considered secret as in [4, 5, 22]—and the case of perfect reconstruction—a scenario of particular relevance for categorical data, where sensitive attributes such as diseases, political opinions, or religious belief are categorical hence partial reconstruction is not immediately meaningful, e.g. [17, 18]. In these settings, we prove closed-form upper bounds that substantially reduce the required noise compared to existing ReRo bounds, and we validate these improvements experimentally.

We leverage our bounds to propose a new RAD-based DP audit strategy. Having a general-purpose auditing tool overcomes the limited scope of previous approaches that focus specifically on detecting bugs [10] or only auditing with respect to specific risk [4]. By grounding the audit in our tight bounds for RAD—covering all possible risks—this approach provides a more accurate and actionable assessment of privacy risks than existing methods. While our auditing framework is general in scope, in this paper we instantiate it for LDP and address key limitations of the state-of-the-art tool, LDP Auditor [4]. Unlike LDP Auditor, which relies on perfect reconstruction without target-specific auxiliary knowledge—and thus misses important threats such as AIAs—our method is both more general and produces *tighter empirical estimates* of the

privacy budget for all the tested LDP mechanisms. In particular, our auditing achieves higher accuracy by overcoming the limitations on privacy budget estimation inherent to LDP AUDITOR as demonstrated in our empirical study (see Figure 6).

Our contributions are summarized as follows:

- We empirically show that ReRo and its existing bounds fail to account for imputation-based success and target-specific auxiliary knowledge, limiting their applicability.
- We introduce the *Reconstruction Advantage (RAD)* as a consistent and unifying risk metric. RAD avoids overestimating risk, and naturally incorporates auxiliary knowledge.
- We establish both a tight worst-case bound under arbitrary auxiliary knowledge and a public-knowledge-dependent, universally tight bound for RAD. Additionally, we derive black-box bounds for attackers lacking target-specific auxiliary knowledge.
- We construct the optimal attack strategy for any reconstruction, mechanism and prior distribution, formally proving its optimality and empirically confirming its utility for auditing and risk assessment.
- We proposed a RAD-based DP auditing framework that provides a broader threat analysis and improved accuracy that existing LDP auditing techniques.

We provide the code used for our experiments accessible in https://github.com/PatriciaBalboaKIT/Understanding-Disclosure-Risk-in-Differential-Privacy.

2. BACKGROUND

In this section, we introduce the relevant concepts for this work and present the notation used throughout the manuscript.

Differential Privacy. We assume each record $z \in \mathcal{Z}$ to be drawn independently from an underlying prior distribution $\mathcal{Z} \sim \pi$. Let $\mathcal{D}(\Theta)$ denote the space of probability distributions over the output space Θ. We consider a learning algorithm $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ which, given an input database $D \in \mathcal{Z}^n$, produces a global output (e.g., an aggregate statistic or a trained model) $\theta = \mathcal{M}_D \in \Theta$ with probability/density function $p_{\mathcal{M}}(\theta \mid D)$. In this context, DP is formalized as follows:

Definition 2.1 ([13])**.** *A mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ is (ε, δ)-differentially private if for all $S \subseteq \Theta$ and for every pair of datasets $D_0, D_1 \in \mathcal{Z}^n$ such that $d_H(D_0, D_1) \leq 1$:*

$$\Pr(\mathcal{M}(D_0) \in S) \leq \mathrm{e}^{\varepsilon} \Pr(\mathcal{M}(D_1) \in S) + \delta$$

where $d_H(D_0, D_1)$ denotes the Hamming distance [38].

If $n = 1$, i.e., $\mathcal{M}$ takes as input a single data record $z \in \mathcal{Z}$, we get the particular case of *Local Differential Privacy* (LDP)–a rigorous and increasingly relevant privacy model in which data is randomized on the client side before being transmitted to a data collector [14], making it especially suitable for privacy-sensitive applications such as telemetry and location-based services where no trusted data curator is considered [16].

The privacy budget ε determines how closely the probabilities of observing the same output on databases D_0 and D_1 must align, hence bounding their statistical "indistinguishability". A smaller ε provides stronger privacy guarantees but typically comes at the cost of utility [14]. The parameter δ allows certain violations of ε-DP while characterizing how likely such failures are to occur and the degree of such failures. Consequently, we aim to parameterize the attack performance based on the privacy parameters.

Differential Privacy and Attack Resilience. Following previous work we consider for any target record z an *informed adversary* [5] with access to: the fixed dataset $D_- = D \backslash \{z\}$, the distribution of data records π, the output θ of the model trained on $D_z = D_- \cup \{z\}$, the mechanism $\mathcal{M}$, and optional target-specific auxiliary knowledge $a(z)$ about target record z.

We adopt this adversary model because, under the assumption that records are independently drawn from π, bounding the performance of such an attacker also bounds the performance of any attacker with less information [5].

Our analysis focuses on DRAs, where the adversary's goal is to correctly reconstruct completely or partially the target record z, potentially given auxiliary knowledge $a(z) \in aux$, about the target. DRAs cover AIAs and MIAs as particular cases [5]: In an MIA, the attacker knows

the entire target record $a(z) = z$ and seeks only to infer its participation in the dataset. In an AIA, records are structured as $z = (x, y)$, $a(z) = x$ is considered public and the attacker aims to perfectly reconstruct the sensitive attribute y. More generally, in a DRA setting, it is natural to assume access to target-specific auxiliary knowledge. For example, when reconstructing a license plate number from a target's car image, the attacker may already know that the car is red or its specific model. Hence, DRAs cover the broad range of commonly discussed privacy risks, including membership inference as a particular instance [5]. Formally, a DRA, denoted by $A\colon \Theta \times aux \to \mathcal{Z}$ takes the output of a DP mechanism $\theta \sim \mathcal{M}(D)$ and the target auxiliary information and produces a candidate $\tilde{z} = A(\theta, a(z), D_-)$. Note that, in case of composing several mechanisms we consider the final output after the whole process.

Then, if the output is similar enough, $\ell(\tilde{z}, z) \leq \eta$, the attack is considered successful. The loss function ℓ depends on the context and use case, for instance, in a classic AIA, given $z = (x, y)$ we define $\phi(z) = y$ and

$$\ell(\tilde{z}, z) = \begin{cases} 0 & \text{if } \phi(\tilde{z}) = \phi(z) \\ 1 & \text{otherwise.} \end{cases}$$

In a MIA, ℓ is the characteristic function such that $\ell(\tilde{z}, z) = 0$ when $\tilde{z} = z$ and one otherwise. However, it may be already sensitive to partially reconstruct the target, for instance, the image domain even if not all pixels are correct, we may gather sensitive information such as the action performed in the image and therefore ℓ is chosen as an image-specific metric, such as the Learned Perceptual Image Patch Similarity (LPIPS) [5]. Given a loss function ℓ and threshold η we define the *success set* of a target z as $S_\eta(z) = \{z' \in \mathcal{Z}\colon \ell(z, z') \leq \eta\}$.

Now the question arises about how to evaluate the performance of a DRA. For the particular cases of AIA and MIAs, the current literature [20, 59] agrees on the following measure:

Definition 2.2 (Adapted from [59])**.** *Given π the distribution of data records and $\mathcal{M}, \phi(z), a(z), A$ as defined above the attribute advantage,* Adv_{AIA}*, is defined as*

$$\Pr_{\substack{z_0 \sim \pi \\ \theta \sim \mathcal{M}(D_{z_0})}} [A(\theta, a(z_0)) = \phi(z_0)] - \Pr_{\substack{z_0, z_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{z_1})}} [A(\theta, a(z_0)) = \phi(z_0)] .$$

The attribute advantage measures the adversary's gain in correctly inferring a sensitive attribute $\phi(z)$ when the record is in the input dataset $z_0 \in D$, compared to when it is drawn from the underlying distribution π. The second term in Theorem 2.2 corrects for cases where the attribute could be inferred even without the record being in the database (e.g., through imputation [28]).

Note that the current proposed performance metric for general DRAs [5] does not define an advantage but instead only accounts for the success probability of an attack that has as input solely the output of the DP mechanism and the known dataset D_-, ignoring any possible target-specific auxiliary knowledge:

Definition 2.3 (ReRo [5])**.** *Let π be a prior over $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ a reconstruction error function. Mechanism $\mathcal{M} : \mathcal{Z}^n \to \mathcal{D}(\Theta)$ is (η, γ)-reconstruction robust with respect to π, ℓ if for any dataset $D_- \in \mathcal{Z}^{n-1}$ and any reconstruction adversary $A : \Theta \to \mathcal{Z}$ it holds that*

$$\Pr_{Z \sim \pi, \theta \sim \mathcal{M}(D_Z)} [\ell(Z, A(\theta)) \leq \eta] \leq \gamma.$$

The first bound for Reconstruction Robustness under ε-DP was given by [5] as:

$$\gamma \leq \kappa^+_{\pi,\ell}(\eta) \mathrm{e}^{\varepsilon}, \tag{1}$$

where $\kappa^+\pi, \ell(\eta) = \sup z_0 \Pr_{Z \sim \pi}[\ell(z_0, Z) \leq \eta]$. Intuitively, $\kappa^+_{\pi,\ell}(\eta)$ represents the success probability of an oblivious attack that always selects the most likely reconstruction under the prior π.

Recent work [22] refined this bound using the f-DP formulation. f-DP [11] is a characterization of the original DP notion that captures the exact statistical indistinguishability between neighbors through the functional f. Formally, it can be defined as follows:

Definition 2.4 ([31]). *Let $f\colon [0,1] \to [0,1]$ be a continuous, convex, non-increasing function such that $f(x) \leq 1 - x$. A mechanism $\mathcal{M}$ satisfies f-DP if, for all pairs of neighboring datasets D_0, D_1 and all binary-valued post-processing algorithms A whose domain contains the range of $\mathcal{M}$,*

$$\Pr(A(\mathcal{M}(D_0)) = 1) \leq 1 - f\big(\Pr(A(\mathcal{M}(D_1)) = 1)\big).$$

Here, f is known as a *trade-off function* [11], named for its interpretation in the context of hypothesis testing. Specifically, consider A as a test of H_0: the input is D_0 vs. H_1: the input is D_1, applied to the output of $\mathcal{M}$. Then $\Pr(A(\mathcal{M}(D_0)) = 1)$ is the significance level and $\Pr(A(\mathcal{M}(D_1)) = 1)$ is the power of the test. Under this interpretation, for a given significance level, f bounds the maximum achievable power. The f-DP framework also facilitates the computation of quantities such as the *total variation* distance between outputs of neighboring datasets.

Definition 2.5. *A mechanism $\mathcal{M}$ has total variation at most $\mathrm{TV}(\mathcal{M})$ if, for all neighboring datasets D_0, D_1,*

$$\sup_{S \subseteq \Theta} |\Pr(\mathcal{M}(D_0) \in S) - \Pr(\mathcal{M}(D_1) \in S)| \leq \mathrm{TV}(\mathcal{M}).$$

For any (ε, δ)-DP mechanism, the TV distance can be bounded [30] as

$$\mathrm{TV}(\mathcal{M}) \leq \max_{\alpha \in [0,1]} \big(1 - f(\alpha) - \alpha\big) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}. \tag{2}$$

Following this formalism, [22] present the following bound for any f-DP mechanism,

$$\gamma \leq 1 - f(\kappa^+_{\pi,\ell}(\eta)). \tag{3}$$

which they showed empirically nearly tight for DP-SGD.

3. Review of the Related Work

In this section, we review the relevant previous work on measuring the effective attack resilience of DP mechanisms for calibration and auditing. We highlight key approaches and discuss novel insights on gaps that motivate our investigation.

Attack-Based DP Noise Calibration. Several recent studies [9, 32] have demonstrated that calibrating the noise of DP mechanisms based on their resilience to specific attacks can significantly help improve utility. Such approaches, however, primarily target MIAs. Tuning the privacy parameters solely to mitigate MIAs may lead to unnecessary utility degradation without offering meaningful privacy benefits in scenarios where membership is public or considered non-sensitive [5].

Beyond MIAs, privacy concerns often involve AIA, where the adversary aims to infer sensitive attributes of individuals from released data [26, 49]. A common metric for evaluating such attacks is the attribute advantage [59]. Existing works that provide theoretical bounds for AIAs typically follow two directions: they either analyze specific attack strategies [59] or adopt more general DRA frameworks [5, 20]. Within the latter, the notion of ReRo has emerged as the metric for measuring the risk of DRAs, under which attribute inference can be modeled as a special case [5]. Moreover, Equation (1) [5] and Equation (3) [22] provide ReRo-based DP noise calibration methods, being the second nearly tight for DP-SGD.

A note on Limitations of ReRo. A general-purpose measure of reconstruction risk would be expected to cover all relevant attack scenarios, including MIAs and AIAs [5]. However, ReRo does not formally account for the impact of target-specific auxiliary knowledge, hence excluding MIAs, AIAs and targeted DRAs as introduced in Section 2.

We note that formally, as specified in Theorem 2.3, the adversary A in the corresponding paper only gets access to mechanism output $\mathcal{M}(D)$, i.e., $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, implying that $\Pr(A(\mathcal{M}(D), z) \in S) = \Pr(A(\mathcal{M}(D), z') \in S)$ for any pair of possible targets z, z' and output set S. Under this assumption, the adversary A cannot adapt its strategy to a specific target z. This choice fundamentally prevents assessing the risk of MIA and AIA, as they do assume full or partial knowledge of some target records. Note that as auxiliary knowledge is often available in reality

and can remarkably increase the risk of attacks. Indeed, most real-world privacy attacks historically exploit publicly available information about the target [42, 45, 52]. We will show in Section 4 that the optimal attack leverages target-specific auxiliary knowledge, and its success highly depends on it.

Not only the original ReRo definition excludes such knowledge, all succeeding formal bounds connecting ReRo and DP were also proven under this restrictive exclusion. The requirement that the attacker depends only on $\mathcal{M}(D)$—ignoring target-specific information—is critical to establishing both Equations (1) and (3). This is not merely a theoretical limitation: we will empirically demonstrate in Section 7 that these bounds fail once we deploy attacks that exploit target-specific knowledge against well-known mechanisms, including DP-SGD.

A direct extension of ReRo to targeted attacks $A(\theta, z)$ fails: Not only do the original bounds no longer hold, but the metric also collapses to a substantial overestimation of risk due to imputation and background knowledge. For instance, the trivial MIA, $A(\theta, z) = z$, has success probability 1, which ReRo would interpret as a catastrophic privacy risk, even though no actual leakage occurs. This is not a negligible edge case; it has caused misleading overestimation of risk in black-box attacks on classification models [28], where much of the reported success arose from data imputation rather than exploiting the mechanism's output. Such overestimation obscures the true leakage and can lead to unnecessary utility loss when ReRo is used to calibrate noise in DP.

Even under the original assumption that the attacker has no target-specific knowledge and A depends only on $\mathcal{M}(D)$, ReRo still overestimates risk, as we discussed in our preliminary work [20]. The mechanism output $\mathcal{M}(D)$ inherently reveals distributional information and population-level statistics, which are the primary goals of any learning process. This information can be used to perform imputation and infer attributes of individual records—even those not in D—with high accuracy, particularly when strong correlations exist (e.g., smoking correlating with cancer). In this case, the apparent attack success is driven by statistical inference rather than actual privacy violations, a phenomenon often referred to as a *privacy fallacy* [12, 31]. Indeed, several works establish that it is impossible to simultaneously provide utility and eliminate absolute information gain [12, 31].

We conclude that ReRo is unreliable as an attack resilience metric, as it overlooks key statistical phenomena that distort privacy risk assessment such as data imputation and targeted attacks. Both cases are very common and have an impact in practice (see Section 7), motivating the need of a novel framework to more accurately assess the risk of DP mechanisms with respect to privacy attacks.

DP Auditing. The tightness of theoretical bounds also directly influences the accuracy of DP auditing outcomes [2, 25, 46]. DP auditing seeks to measure or estimate the effective privacy guarantees of a system, often in a black-box or post-deployment setting. Particularly, literature on auditing demonstrates tight estimates of the privacy budget [46], discovers implementation flaws [54], and estimates empirical privacy [25]. However, most existing auditing approaches concentrate on MIAs [2, 25, 43, 51], which limits their ability to detect broader forms of privacy leakage.

Some auditing techniques extend beyond MIAs to consider AIAs, but these are restricted to specific contexts—such as Label DP [40] or synthetic data generation [23]. In the local differential privacy (LDP) setting, the state-of-the-art framework LDP Auditor [4] relies specifically on perfect reconstruction without target-specific auxiliary knowledge for auditing.

We conclude that there are no reliable adversarial bounds on the advantage of attacks beyond MIAs, nor for AIAs or DRAs. In our preliminary work [20], we provide intermediate solutions for attackers who only use the mechanism output; however, the impact of target-specific auxiliary information remains unexplored. This is not something we can ignore: auxiliary information about individuals has been a core component of attacks showing privacy risk. Particularly, the classical census attack [52] could not be possible without the (auxiliary) knowledge of the demographic public attributes. Moreover, to date, there is no general auditing framework grounded in DRAs that can systematically evaluate privacy leakage across diverse DP mechanisms.

4. Reconstruction Advantage

In this section, we introduce reconstruction advantage (RAD) as a novel, unifying metric for adversarial risk assessment. We first establish a worst-case bound on RAD that holds for any mechanism, data distribution, and auxiliary knowledge, ensuring robustness when the attacker's prior knowledge is unknown. We then refine this result by deriving a tighter bound under known auxiliary knowledge and by constructing the corresponding optimal attack that achieves it. Together, these results demonstrate that our bounds are tight: the worst-case bound applies when auxiliary knowledge is uncertain, while the refined bound is exact when such knowledge is specified. Finally, we empirically validate our optimal attack on real datasets in both private learning and LDP settings, confirming the practical tightness of our theoretical estimations (see Section 7.3).

In order to address ReRo's lack of accounting for the impact of target-specific auxiliary knowledge, we explicitly incorporate this concept into our definition. Formally, each record $z \in \mathcal{Z}$ may be associated with target-specific auxiliary information $a(z) \in aux$. The auxiliary information can take different forms. For instance, in the classical attribute inference attack (AIA) setting, where records are pairs $z = (x, y)$, one may define $a(z) = x$ and attempt to infer y. Alternatively, in the image reconstruction setting, the target may be the full record z, while $a(z)$ could correspond to a label such as "image of a person" or "image of an animal". The only structural assumption we impose is that the type of auxiliary information is consistent across all records: if $a(z)$ corresponds to a set of pixels, then for any other record z', $a(z')$ must also be a set of pixels (and not, for example, a semantic label).

Having established this formalization, we are now in a position to introduce our metric[1].

Definition 4.1 (η-RAD)**.** *Let π be a prior over $\mathcal{Z}$, $\ell\colon \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ an error function, and $a(z) \in aux$ the target-specific auxiliary information for each $z \in \mathcal{Z}$. Given a mechanism $\mathcal{M} : \mathcal{Z}^n \to \mathcal{D}(\Theta)$, any dataset $D_- \in \mathcal{Z}^{n-1}$ and any adversary $A : \Theta \times aux \to \mathcal{D}(\mathcal{Z})$ we define the η-reconstruction advantage, η-RAD, as*

$$\Pr_{\substack{Z_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{Z_1})}} [\ell(Z_1, A(\theta, a(Z_1))) \leq \eta] - \Pr_{\substack{Z_0, Z_1 \sim \pi \\ \theta \sim \mathcal{M}(D_{Z_0})}} [\ell(Z_1, A(\theta, a(Z_1)) \leq \eta].$$

This definition explicitly accounts for target-specific auxiliary knowledge, providing a generalization of the membership and attribute advantages to arbitrary reconstruction attacks. η-RAD takes values between -1 and $(1 - \kappa_\pi) \leq 1$ where $\kappa_\pi = \Pr_{Z,Z' \sim \pi}[Z = Z']$, i.e. the probability of resampling from the distribution π, analogous to the cases of membership and attribute advantage [59]. Intuitively, RAD measures the increase in the attacker's success probability that arises solely from the target's participation in the private learning process. In this way, RAD avoids the overestimation of risk that is inherent in ReRo. If RAD ≤ 0, participation carries no risk, since the attacker's probability of correctly reconstructing the record is no greater than if the individual had not participated. Larger values of RAD indicate higher participation risk. In the extreme case where RAD $= 1 - \kappa_\pi$, participation entails absolute risk: the attacker always succeeds in reconstructing the participant's record, while no sensitive information can be reconstructed from non-participants.

Previous bounds for ReRo assume that reconstruction attacks perform equally for every target. This assumption holds when the adversary has no target-specific auxiliary knowledge, but breaks once *aux* is available: for instance, knowing that a target's surname is "Smith" might give less information than knowing that it is "Sainthorpe-Burton", as the latter is less frequent and hence carries more information. Such differences are not captured by ReRo, nor reflected in the proofs of the corresponding bounds [5, 20], which consequently fail for attacks utilizing target-specific auxiliary knowledge as revealed in Section 7.3. Hence, we provide the first theoretical bound that explicitly accounts for *aux* and covers any possible attack from MIAs to the most general DRAs:

[1]Note that we presented a preliminary idea for this metric in [20], initially calling it U-ReRo, which, however, similar to ReRo, fails to take *aux* into account.

Theorem 4.2 ((ε,δ)-DP implies η-RAD)**.** *Let $\pi, \ell, \eta \geq 0$ as in Def. 4.1, and $\kappa_\pi = \Pr_{Z,Z'\sim\pi}[Z = Z']$. If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies (ε,δ)-DP, then for any attack $A\colon \Theta\times aux \to \mathcal{D}(\mathcal{Z})$, and database D_- we have*

$$\eta\text{-RAD} \leq \mathrm{TV}(\mathcal{M})(1-\kappa_\pi) \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}(1-\kappa_\pi).$$

Proof. We use $\int f(x)\,\mathrm{d}\mu(x)$ as unified notation that represents either a sum (if μ is the counting measure) or an integral (if μ is the Lebesgue measure), hence aggregating both the discrete and continuous case in one. First, note that that for every $z \in \mathcal{Z}$ and target-specific knowledge $a(z)$, any attack defines $A(D, a(z)) \equiv \mathcal{A}_z(\mathcal{M}(D))$ verifying

$$p_{\mathcal{A}_z}(s \mid D) \equiv p_A(s \mid a(z), D) = \int_\Theta p_{\mathcal{M}}(\theta \mid D)\, p_A(s \mid \theta, a(z))\,\mathrm{d}\mu(\theta).$$

Therefore,

$$\begin{aligned}
\mathrm{TV}(\mathcal{A}_z(D), \mathcal{A}_z(D')) &\coloneqq \sup_S \left|\Pr(\mathcal{A}_z(D) \in S) - \Pr(\mathcal{A}_z(D') \in S)\right| && (4)\\
&= \tfrac{1}{2}\int_{\mathcal{Z}} \left| p_A(s \mid \mathcal{M}(D), a(z)) - p_A(s \mid \mathcal{M}(D'), a(z))\right| \mathrm{d}\mu(s) && (5)\\
&= \tfrac{1}{2}\int_{\mathcal{Z}} \left| \int_\Theta p_A(s \mid \theta, a(z))\,(p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D'))\,\mathrm{d}\mu(\theta)\right| \mathrm{d}\mu(s) && (6)\\
&\leq \tfrac{1}{2}\int_{\mathcal{Z}} \int_\Theta p_A(s \mid \theta, a(z))\,|p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D')|\,\mathrm{d}\mu(\theta)\,\Big|\,\mathrm{d}\mu(s) && (7)\\
&= \tfrac{1}{2}\int_\Theta |p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D')|\,\mathrm{d}\mu(\theta) \int_{\mathcal{Z}} p_A(s \mid \theta, a(z))\,\mathrm{d}\mu(s) && (8)\\
&= \tfrac{1}{2}\int_\theta |p_{\mathcal{M}}(\theta \mid D) - p_{\mathcal{M}}(\theta \mid D')|\,\mathrm{d}\mu(\theta) && (9)\\
&= \mathrm{TV}(\mathcal{M}(D), \mathcal{M}(D')), && (10)
\end{aligned}$$

where Equation (5) follows from [36, Proposition 4.2, p. 48] and Equation (7) from Minkowski's inequality.

Note that given any success set $S_\eta(z) = \{\theta \in \Theta\colon \ell(z,\theta) \leq \eta\}$, using $A(D, a(z)) \equiv \mathcal{A}_z(\mathcal{M}(D))$ notation we have

$$\Pr_{\substack{Z_1\sim\pi\\ \theta\sim\mathcal{M}(D_{Z_0})}}[\ell(Z_1, A(\theta, a(Z_1))) \leq \eta] = \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)].$$

Hence, applying Equation (10) and Theorem 2.5 to RAD Theorem 4.1 we obtain:

$$\begin{aligned}
\eta\text{-RAD} &= \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_0,Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\\
&= \mathop{\mathbb{E}}_{Z_0\sim\pi}\left[\Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\right]\\
&= \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}(\Pr[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)])\right]\\
&\overset{10}{\leq} \mathrm{TV}(\mathcal{M}) \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right].
\end{aligned}$$

Since, $\mathbb{E}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right] = 1 - \sum_z \pi_z^2$ for discrete variables and 1 for continuous ones, it follows that η-RAD$\leq \mathrm{TV}(\mathcal{M})$. Finally, Equation (2) completes the result. □

Note that when $\mathcal{Z}$ countable $\kappa_\pi = \sum_z \pi_z^2$ and in the uncountable case the result simplifies to η-RAD$\leq \mathrm{TV}(\mathcal{M})$.

Theorem 4.2 is the first general bound for η-RAD under the strongest adversarial model, where the attacker may use auxiliary knowledge. Experiments on real datasets provide supporting evidence that this bound is tight (see Section 7): attacks can achieve the predicted advantage, confirming that it accurately captures the worst-case scenario. Hence, becoming a crucial tool for effective DP noise calibration while overcoming the limitations of ReRo.

Theorem 4.2 does not depend on the attacker's auxiliary knowledge. In particular, the same guarantee holds whether the attacker has no auxiliary information ($aux = \varnothing$) or complete knowledge of the record ($a(z) = z$), since the result is derived in a worst-case manner. However, when the attacker's goal is to reconstruct an entire record (as in DRA) or infer parts of it (as in AIA), it is unreasonable to assume that the attacker already knows the full record ($a(z) = z$)—an assumption more appropriate for MIA. Therefore, we next refine our analysis and develop a tighter bound that explicitly incorporates the attacker's target-specific auxiliary knowledge.

Theorem 4.3. *Let $\pi, \ell, \eta \geq 0$ as in Def. 4.1. Given $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$, then for any attack $A\colon \Theta \times aux \to \mathcal{D}(\mathcal{Z})$, database D_-, and η such that $S_\eta(z) \cap S_\eta(z') = \varnothing$ for all $a(z) = a(z')$, we have*[2]

$$\eta\text{-RAD} \leq \sum_{\theta\in\Theta} \sum_{x\in aux} \max_{\substack{a(z)=x \\ w(z,\theta)>0}} w(z,\theta)\pi_z,$$

where $w(z,\theta) = p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)$.

Proof. We denote by μ the counting measure in the discrete case and the Lebesgue measure in the continuous case. First, using probability properties, we rewrite RAD definition as

$$\begin{aligned}
\eta\text{-RAD} &= \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_0,Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)] \\
&= \mathop{\mathbb{E}}_{Z_0\sim\pi}\Big[\Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\Big] \\
&= \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\Big[\mathbf{1}_{\{Z_0\neq Z_1\}}\big(\Pr[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\big)\Big] \\
&= \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\Big[\mathbf{1}_{\{Z_0\neq Z_1\}}\int_\Theta p_A(S_\eta(Z_1) \mid \theta, a(Z_1))\Big(p_{\mathcal{M}}(\theta \mid D_{Z_1}) - p_{\mathcal{M}}(\theta \mid D_{Z_0})\Big)\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{Z_1\sim\pi}\Big[\int_\Theta p_A(S_\eta(Z_1) \mid \theta, a(Z_1)) \mathop{\mathbb{E}}_{Z_0\sim\pi}\Big[\mathbf{1}_{\{Z_0\neq Z_1\}}(p_{\mathcal{M}}(\theta \mid D_{Z_1}) - p_{\mathcal{M}}(\theta \mid D_{Z_0}))\Big]\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{Z_1\sim\pi}\Big[\int_\Theta p_A(S_\eta(Z_1) \mid \theta, a(Z_1))\Big(p_{\mathcal{M}}(\theta \mid D_{Z_1}) \mathop{\mathbb{E}}_{Z_0\sim\pi}[\mathbf{1}_{\{Z_0\neq Z_1\}}] - \mathop{\mathbb{E}}_{Z_0\sim\pi}[\mathbf{1}_{\{Z_0\neq Z_1\}}p_{\mathcal{M}}(\theta \mid D_{Z_0})]\Big)\,\mathrm{d}\mu(\theta)\Big] \\
&= \mathop{\mathbb{E}}_{Z_1\sim\pi}\Big[\int_\Theta p_A(S_\eta(Z_1) \mid \theta, a(Z_1)) \underbrace{\big(p_{\mathcal{M}}(\theta \mid D_{Z_1}) - p_{\mathcal{M}}(\theta)\big)}_{w(z_1,\theta)}\,\mathrm{d}\mu(\theta)\Big]. \qquad (11)
\end{aligned}$$

Where Equation (11) follows trivially for the continuous case since $\mathbb{E}_{Z_0\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right] = 1$ and for the discrete one since

$$\begin{aligned}
& p_{\mathcal{M}}(\theta \mid D_{Z_1}) \mathop{\mathbb{E}}_{Z_0\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right] - \mathop{\mathbb{E}}_{Z_0\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}p_{\mathcal{M}}(\theta \mid D_{Z_0})\right] && (12) \\
&= p_{\mathcal{M}}(\theta \mid D_{Z_1})(1-\pi_1) - \mathop{\mathbb{E}}_{Z_0\sim\pi}[p_{\mathcal{M}}(\theta \mid D_{Z_0})] + p_{\mathcal{M}}(\theta \mid D_{Z_1})\pi_1 && (13) \\
&= p_{\mathcal{M}}(\theta \mid D_{Z_1}) - p_{\mathcal{M}}(\theta). && (14)
\end{aligned}$$

Now, for all z_1, z_2 such that $a(z_1) = a(z_2) = x$, and for any fixed output θ, we have that

$$\Pr_A(S_\eta(z_1) \mid \theta, a(z_1)) = \Pr_A(S_\eta(z_1) \mid \theta, a(z_2)) = \Pr_A(S_\eta(z_1) \mid \theta, x).$$

Hence, given $a^{-1}(x) = \{z\colon a(z) = x\}$ for all $x \in aux$, we obtain

$$\begin{aligned}
\text{RAD} &= \int_{\mathcal{Z}}\int_\Theta \Pr_A(S_\eta(z) \mid a(z), \theta)w(z,\theta)\pi_z\,\mathrm{d}\mu(z)\,\mathrm{d}\mu(\theta) \\
&= \int_\Theta\int_{aux}\int_{a^{-1}(x)} \Pr_A(S_\eta(z) \mid x, \theta)w(z,\theta)\pi_z\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(z)\,\mathrm{d}\mu(\theta) \\
&\leq \int_\Theta\int_{aux}\int_{a^{-1}(x)} \Pr_A(S_\eta(z) \mid x, \theta)[w(z,\theta)]_{\geq 0}\pi_z\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(z)\,\mathrm{d}\mu(\theta)
\end{aligned}$$

[2]Previous sums must be changed by integrals for continuous variables.

$$\leq \int_{\Theta} \int_{aux} \max_{\substack{z\colon a(z)=x \\ w(z,\theta)>0}} w(z,\theta)\pi_z \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(\theta), \tag{15}$$

where Equation (15) holds since by hypothesis $S_\eta(z)$ are mutually disjoint for all $z \in a^{-1}(x)$, therefore

$$\int_{a^{-1}(x)} \Pr_A(S_\eta(z_1) \mid x, \theta)\,\mathrm{d}\mu(z_1) \leq 1. \qquad \square$$

Theorem 4.3 is a white-box bound that can be computed when the specific $\mathcal{M}$ and auxiliary knowledge, aux, is known. At the same time, it becomes more precise than our black-box bound Theorem 4.2. For instance, in a MIA, i.e., $a(z) = z$ for all records and $\ell(z, z') = 0 \Leftrightarrow z = z'$, we get

$$\mathrm{Adv}_{MIA} \equiv \text{0-RAD} \leq \sum_{z} \sum_{\theta\colon w(\theta,z)>0} w(\theta, z)\pi_z. \tag{16}$$

On the other extreme, if $aux = \varnothing$ we obtain

$$\text{0-RAD} \leq \sum_{\theta} \max_{z} w(\theta, z)\pi_z. \tag{17}$$

Given $|\mathcal{Z}| = m$, previous equation admits the simplification

$$\text{0-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) \right) \pi_z, \tag{18}$$

where $\Theta_1 = \{\theta \in \Theta\colon z_1 \in \arg\max_j p_{\mathcal{M}}(\theta \mid z_j)\}$ and for every $i \geq 1$, Θ_{i+1} is recursively defined as

$$\Theta_{i+1} = \{\theta \in \Theta\colon z_{i+1} \in \arg\max_j p_{\mathcal{M}}(\theta \mid z_j)\} \setminus \cup_{k=1}^{i} \Theta_k.$$

Note that Theorem 4.3 assumes $S_\eta(z) \cap S_\eta(z') = \varnothing$ for all $a(z) = a(z')$. This assumption is well-motivated by the semantics of the attack models we consider. Take first the case of attribute inference (AIA), where $\eta = 0$ and $S_\eta(z) = \{z'\colon \phi(z') = \phi(z) = y\}$, i.e., the set of records sharing the same sensitive attribute. If $a(z) = a(z')$ and $S_\eta(z) \cap S_\eta(z') \neq \varnothing$, then we must have $\phi(z) = \phi(z') = y$. Since $z = (x, y)$ and $z' = (x, y)$ share both the auxiliary attribute x and the sensitive attribute y, it follows that $z = z'$, consistent with our assumption. Similarly, in the case of perfect reconstruction (as in the original ReRo setting [22]), where $S_0(z) = \{z\}$, we have $S_0(z) \cap S_0(z') \neq \varnothing$ if and only if $z = z'$, regardless of the auxiliary knowledge. In summary, the intuition behind this assumption is that, if two individuals with the same public information also share the same sensitive information, then the auxiliary knowledge alone suffices to extract the private information, hence the attacker does not need to rely on the mechanism's output.

We illustrate the benefit of Theorem 4.3 computing our bound for relevant DP mechanisms in the following examples and visualize them in Figure 3.

Example 4.4 (GRR)**.** *The generalized randomized response mechanism (GRR) [29] its a LDP mechanism that outputs the true record z_1 with probability $p = \mathrm{e}^\varepsilon/(\mathrm{e}^\varepsilon + m - 1)$ and any other record $z_0 \neq z_1$ with probability $q = (\mathrm{e}^\varepsilon + m - 1)^{-1}$, hence*

$$w(\theta, z) = \begin{cases} (p - q)(1 - \pi_\theta) & \text{if } z = \theta \\ (q - p)\pi_\theta & \text{otherwise.} \end{cases} \tag{19}$$

Since $q - p \leq 0$, we have that $w(z, \theta) > 0$ iff $z = \theta$. Hence, applying Theorem 4.3 we obtain

$$\eta\text{-RAD} = \sum_{\theta \in \mathcal{Z}} \sum_{z \in \mathcal{Z}} (p - q)(1 - \pi_z)\pi_z = \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + m - 1}(1 - \sum_z \pi_z^2) = \mathrm{TV}(\mathrm{GRR})(1 - \kappa_\pi). \tag{20}$$

Example 4.5 (OUE and $aux = \varnothing$)**.** *In the optimal unary encoding (OUE) mechanism [55] each user's input $z \in \mathcal{Z}$ as a one-hot m-dimensional binary vector and perturbs each bit independently. For each position $i \in [m]$, the obfuscated vector θ is sampled such that $\Pr[\theta_i = 1] = 1/2$ if $i = z$,*

and $q = \frac{1}{e^{\varepsilon}+1}$ *otherwise. Denoting* $p = 1-q$ *and* $k_\theta = \#\{\theta_i = 1\}$*, we have that every* θ *such that* $k_\theta \geq 1$

$$\Pr(\theta, z) = \begin{cases} \frac{1}{2}q^{k_\theta - 1}p^{m-k_\theta} & \text{if } \theta_z = 1 \\ \frac{1}{2}q^{k}p^{m-k_\theta - 1} & \text{if } \theta_z \neq 1 \end{cases} \tag{21}$$

and $\Pr(\vec{0}, z) = \frac{1}{2}p^{m-1}$*. Since there are* $\binom{m}{k}$ *different vectors* θ*, with* $k_\theta = k$*,*

$$\sum_{\theta \in \{0,1\}^m} \max_z \Pr_{\mathcal{M}}(\theta \mid z) = \sum_{\theta \neq \vec{0}} \frac{1}{2}q^{k_\theta - 1}p^{m-k_\theta} + \frac{1}{2}p^{m-1} \tag{22}$$

$$= \frac{1}{2}\sum_{\theta \in \{0,1\}^m} q^{k_\theta - 1}p^{m-k_\theta} = \frac{1}{2}\left(\sum_{k=1}^{m}\binom{m}{k}p^{m-k}(1-p)^{k-1} + p^{m-1}\right) \tag{23}$$

$$= \frac{1}{2}\left(\frac{1-p^m}{1-p} + p^{m-1}\right) = \frac{(1 + p^{m-1} - 2p^m)}{2(1-p)}. \tag{24}$$

Hence, according to Theorem 4.3, we obtain that for $\pi = U[m]$

$$0\text{-RAD} \leq \frac{1}{m}\left(\frac{(1 + p^{m-1} - 2p^m)}{2(1-p)} - 1\right) = \frac{(2p-1)\left(1-p^{m-1}\right)}{2m(1-p)} = \frac{e^{\varepsilon}-1}{2m}\left(1 - \left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right)^{(m-1)}\right).$$

Note that

$$\lim_{\varepsilon \to \infty} \frac{e^{\varepsilon}-1}{2m}\left(1 - \left(\frac{e^{\varepsilon}}{1+e^{\varepsilon}}\right)^{(m-1)}\right) = \frac{m-1}{2m},$$

hence even if we keep reducing the noise (increasing ε*), the attacker's advantage is limited.*

Example 4.6 (SS, $aux = \varnothing$)**.** *In the subset selection mechanism (SS) [58] users report a subset* $\theta \subseteq \mathcal{Z} = \{z_1, \ldots, z_m\}$ *containing their true value* z *with probability* $p = \frac{\omega e^{\varepsilon}}{\omega e^{\varepsilon} + m - \omega}$*, where* $\omega = |\theta| = \max\left(1, \left\lfloor \frac{m}{e^{\varepsilon}+1} \right\rfloor\right)$*. The subset is completed by sampling uniformly from* $\mathcal{Z} \setminus \{z\}$*.*

Note that, given $A = \binom{m-1}{\omega-1}$ *and* $B = \binom{m-1}{\omega}$*,*

$$\Pr_{\mathcal{M}}(\theta \mid z) = \begin{cases} \frac{p}{A} & \text{if } z \in \theta \\ \frac{1-p}{B} & \text{if } z \notin \theta \end{cases} \tag{25}$$

Since $|\Theta| = \binom{m}{\omega}$ *we have that, according to Theorem 4.3, for* $\pi = U[m]$*,*

$$0\text{-}RAD \leq \frac{1}{m}\sum_{\theta \in \Theta} \max_z p_{\mathcal{M}}(\theta \mid z) - 1) \tag{26}$$

$$= \frac{1}{m}\binom{m}{\omega}\frac{p}{A} = \frac{m}{m\omega}p - \frac{1}{m} = \frac{pm - \omega}{m\omega}. \tag{27}$$

Example 4.7 (Gaussian mechanism and $aux = \varnothing$)**.** *The Gaussian mechanism adds Gaussian noise* $\mathcal{N}(0, \sigma)$ *the query value* $q(D) \in \mathbb{R}$ *[6]. If* $\mathcal{Z} = \{z_1, \ldots, z_m\}$ *uniformly distributed and* $\Delta q = 1$*, applying Equation (18) we obtain*

$$0\text{-}RAD \leq \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - p_{\mathcal{M}}(\Theta_i)\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - 1\right) \tag{28}$$

Note that for each z*,* $\Pr_{\mathcal{M}}(\theta \mid z) = \Pr_{M}(\theta \mid q(D_z))$*. Since* D_- *is fixed,* $q(D_z)$ *is completely determined by* z*, hence we use the abuse of notation* $q(D_z) \equiv z$*. We want to compute* $\Pr_{\mathcal{M}}(\Theta_i \mid z_i)$ *for* $i \in [m]$*. Without loss of generality we re-order* $z_1 < z_2 < \cdots < z_n$*, and define the gaps* $\Delta_i := z_{i+1} - z_i$*. For fixed* θ*, the maximizing density corresponds to the* z_i *closest to* θ*. Thus* $\mathbb{R}$ *is partitioned into Voronoi intervals:*

$$\Theta_1 = \left(-\infty, \tfrac{z_1+z_2}{2}\right], \tag{29}$$

$$\Theta_i = \left[\tfrac{z_{i-1}+z_i}{2}, \tfrac{z_i+z_{i+1}}{2}\right], \quad 2 \leq i \leq n-1, \tag{30}$$

$$\Theta_n = \left[\tfrac{z_{n-1}+z_n}{2}, \infty\right). \tag{31}$$

On Θ_i, the maximizer is z_i. Hence Let Φ denote the standard normal CDF and φ its density function. Then, for $i = 1$

$$\Pr_{\mathcal{M}}(\Theta_1 \mid z_1) = \int_{\Theta_1} \varphi_\sigma(\theta - z_1)\, \mathrm{d}\theta = \Phi\Big(\tfrac{(z_1+z_2)/2-z_1}{\sigma}\Big) = \Phi\Big(\tfrac{\Delta_1}{2\sigma}\Big).$$

For $i = m$

$$\Pr_{\mathcal{M}}(\Theta_m \mid z_m) = \int_{\Theta_m} \varphi_\sigma(\theta - z_m)\, \mathrm{d}\theta = 1 - \Phi\Big(\tfrac{(z_{m-1}+z_m)/2-z_m}{\sigma}\Big) = \Phi\Big(\tfrac{\Delta_{n-1}}{2\sigma}\Big).$$

Finally, for $2 \le i \le m-1$,

$$\begin{aligned}\Pr_{\mathcal{M}}(\Theta_i \mid z_i) = \int_{\Theta_i} \varphi_\sigma(\theta - z_i)\, \mathrm{d}\theta &= \Phi\Big(\tfrac{(z_i+z_{i+1})/2-z_i}{\sigma}\Big) - \Phi\Big(\tfrac{(z_{i-1}+z_i)/2-z_i}{\sigma}\Big) \\ &= \Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) - \Phi\Big(-\tfrac{\Delta_{i-1}}{2\sigma}\Big) \\ &= \Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) + \Phi\Big(\tfrac{\Delta_{i-1}}{2\sigma}\Big) - 1,\end{aligned}$$

using $\Phi(-x) = 1 - \Phi(x)$. Therefore

$$\begin{aligned}\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) &= \Phi\Big(\tfrac{\Delta_1}{2\sigma}\Big) + \sum_{i=2}^{m-1} \Big(\Phi\Big(\tfrac{\Delta_i}{2\sigma}\Big) + \Phi\Big(\tfrac{\Delta_{i-1}}{2\sigma}\Big) - 1\Big) + \Phi\Big(\tfrac{\Delta_{m-1}}{2\sigma}\Big) \\ &= 2\sum_{j=1}^{m-1} \Phi\Big(\frac{\Delta_j}{2\sigma}\Big) \;-\; (m-2),\end{aligned}$$

since each Δ_j appears exactly twice in the sum (once from its left neighbor, once from its right). Hence,

$$0\text{-}RAD \le \frac{2}{m}\sum_{j=1}^{m-1} \Phi\Big(\frac{\Delta_j}{2\sigma}\Big) \;- \frac{m-1}{m} \tag{32}$$

$$\le \frac{2(m-1)}{m}\Phi\Big(\frac{1}{(m-1)}\sum_{j=1}^{m-1}\frac{\Delta_j}{2\sigma}\Big) \;- \frac{m-1}{m} \tag{33}$$

$$\le \frac{m-1}{m}\left(2\Phi\Big(\frac{1}{2\sigma(m-1)}\Big) \;- 1\right). \tag{34}$$

Where, Equation (33) follows since $\Delta_j \ge 0$, hence Φ concave, and we can apply Jensen's inequality, and Equation (34) since $\Delta q = 1$ therefore, $\sum_{j=1}^{m-1} \Delta_j = \Delta q = 1$.

Example 4.8 (Laplace Mechanism and $aux = \varnothing$)**.** *The Laplace mechanism adds Laplace noise with scale $b = \Delta q/\varepsilon$ to the query value $q(D) \in \mathbb{R}$ [14]. If $\mathcal{Z} = \{z_1, \ldots, z_m\}$ uniformly distributed and $\Delta q = 1$, analogously to Theorem 4.7, applying Equation (18) we obtain*

$$0\text{-}RAD \le \frac{1}{m}\sum_{i=1}^{m}\Big(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - p_{\mathcal{M}}(\Theta_i)\Big) = \frac{1}{m}\sum_{i=1}^{m}\Big(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - 1\Big) \tag{35}$$

Analogously to the Gaussian case, we use the abuse of notation $z \equiv q(D_z)$. We want to compute $\Pr_{\mathcal{M}}(\Theta_i \mid z_i)$ for $i \in [m]$. Without loss of generality we re-order $z_1 < z_2 < \cdots < z_n$, and define the gaps $\Delta_i := z_{i+1} - z_i$. For fixed θ, the maximizing density corresponds to the z_i closest to θ. Thus $\mathbb{R}$ is again partitioned into Voronoi intervals from Theorem 4.7. Given the Laplace distribution CDF

$$F_i(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-z_i}{b}\right) & \text{if } x < z_i \\ 1 - \frac{1}{2}\exp\left(-\frac{x-z_i}{b}\right) & \text{if } x \ge z_i \end{cases}, \tag{36}$$

for $i = 1$,

$$\Pr_{\mathcal{M}}(\Theta_1 \mid z_1) = F(\frac{z_1+z_2}{2}) - F(-\infty) = 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_1}{2}\right),$$

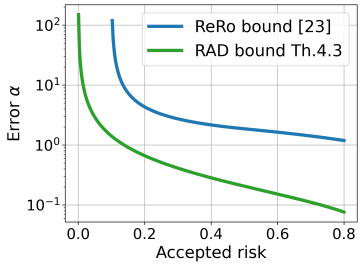


FIGURE 1. Laplace, $\Delta = 1$

FIGURE 2. Upper bound on the mechanism error at 95% confidence when the noise is calibrated using ReRo vs. RAD. We see that for the same risk estimation, calibrating with using RAD yields notable error reduction.

for $i = m$,

$$\Pr_{\mathcal{M}}(\Theta_m \mid z_m) = 1 - F(\frac{z_m + z_{m-1}}{2}) = 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_{m-1}}{2}\right),$$

and for the reminder $2 \leq i < m$*:*

$$\Pr_{\mathcal{M}}(\Theta_m \mid z_m) = F(\frac{z_i + z_{i+1}}{2}) - F(\frac{z_{i-1} + z_i}{2}) = 1 - \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_i}{2}\right) + \frac{1}{2}\exp\left(-\varepsilon\frac{\Delta_{i-1}}{2}\right),$$

Hence,

$$\sum_{i=1}^{m}\Pr_{\mathcal{M}}(\Theta_i \mid z_i) = m - \frac{1}{2}\left(\mathrm{e}^{-\frac{\varepsilon\Delta_1}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{m-1}}{2}} + \sum_{i=2}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_i}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{i-1}}{2}}\right) = m - \sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}} \tag{37}$$

since each Δ_j *appears exactly twice in the sum (once from its left neighbor, once from its right). Hence,*

$$0\text{-}RAD \leq \frac{1}{m}\left(m - 1 - \sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}}\right) \tag{38}$$

$$\leq \frac{m-1}{m} - \frac{1}{m}\sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}} \tag{39}$$

$$\leq \frac{m-1}{m} - \frac{m-1}{m}\mathrm{e}^{-\frac{1}{m-1}\sum_j\frac{\varepsilon\Delta_j}{2}} \tag{40}$$

$$\leq \frac{m-1}{m}\left(1 - \mathrm{e}^{-\frac{\varepsilon}{2(m-1)}}\right). \tag{41}$$

Where, Equation (40) *follows since* $\Delta_j \geq 0$*, hence* Φ *concave, and we can apply Jensen's inequality, and Equation* (41) *since* $\Delta q = 1$ *therefore,* $\sum_{j=1}^{m-1}\Delta_j = \Delta q = 1$.

Previous examples show the applicability of Theorem 4.3 to estimate the risk in real-world scenarios. In Figure 3 we see the improvement when we target specific auxiliary knowledge instead of using our black-box bound (Theorem 4.2). Hence, Theorem 4.3 offers an improved noise calibration method to ensure protection against real attacks, especially when the auxiliary knowledge is well defined. For instance, when the entire record is considered private, we assume that no information is available ($aux = \varnothing$); alternatively, when a specific attribute y is deemed sensitive, we consider all the remainder record public, i.e., $a(z) = z \backslash y$.

Importantly, both Theorems 4.2 and 4.3 address the previous overestimation of ReRo risk, thereby significantly improving the accuracy of the mechanisms when used for noise calibration. In Figure 2, we illustrate the utility gain of our RAD bounds compared to the best existing ReRo bound [22] for the Laplace mechanism. Specifically, we consider $aux = \varnothing$–allowing comparison with [22]. We plot the upper bound on the query error that can be guaranteed with 95%

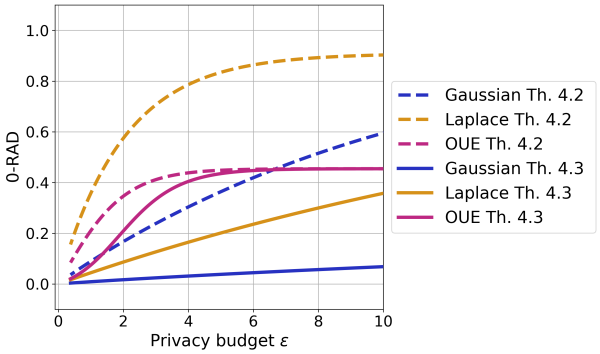


FIGURE 3. Improvement of Theorem 4.3 over Theorem 4.2 for different DP mechanisms, $|\mathcal{Z}| = m$ and uniform prior.

Algorithm 1: Optimal Attack

Input : θ and $a(z) = x$
Output : $\tilde{z} = z^*$
for $z\colon a(z) = x$ **do**
  compute $w(\theta, z)$;
Select $z^* \in \arg\max_{z\colon a(z)=x} w(\theta, z)\pi_z$
if $w(\theta, z^*) > 0$ **then**
  $\tilde{z} = z^*$;
else
  $\tilde{z} \longleftarrow U[\mathcal{Z} \backslash \{z\colon a(z) = x\}]$;

confidence, for $|Z| = 10$ and $\Delta = 1$, showing a substantial improvement in utility enabled by our novel bounds.

Crucially, the Theorem 4.3 bound is universally tight: Given any mechanism and auxiliary knowledge, there exists an attack that attains the bound, hence it can not be further improved. We demonstrate this statement by constructing such an attack explicitly in Algorithm 1, thereby proving the existence of an optimal adversary for a given auxiliary model. In the extreme informed case–when the adversary knows the entire target record (as in an informed MIA)–the optimal strategy is to declare the target a member whenever the mechanism's output provides any positive evidence of participation, i.e., whenever $w(\theta, z) > 0$. Intuitively, if the posterior weight on the true record increases at all after observing θ, the attacker should assume membership. If there is more than one z such that $a(z) = x$, the attacker can not optimize for all at the same time, therefore takes the one maximizing the posterior weight as long as it provides positive evidence. In the extreme case, when $aux = \varnothing$ (no auxiliary information), the attacker cannot narrow the candidate set and the optimal reconstruction picks $z^* \in \arg\max_{z\in\mathcal{Z}} w(z, \theta)\pi(z)$ i.e., any record that maximizes the posterior weight given θ. Applying Theorem 4.3 we can prove that this attack is optimal:

Corollary 4.9 (Attack Optimality)**.** *Given the conditions as in Theorem 4.3, Algorithm 1 achieves the highest attainable η-RAD.*

Proof. Since $S_\eta(z) \cap S_\eta(z') = \varnothing$ for all $a(z) = a(z')$, Algorithm 1 satisfies:

$$\Pr_A(S_\eta(z_1) \mid a(z_1), \theta) = \Pr_A(z_1 \mid a(z_1), \theta) = \mathbf{1}_{\{z_1 = z^* \wedge w(z^*, \theta) > 0\}} \tag{42}$$

When substituting in RAD, using the reformulation as in Equation (11), we get

$$\begin{aligned} \eta\text{-RAD} &= \mathop{\mathbb{E}}_{Z_1 \sim \pi}\Big[\int_\Theta p_A(S_\eta(Z_1) \mid \theta, a(Z_1))\, w(z_1, \theta)\, \mathrm{d}\mu(\theta)\Big] \\ &= \mathop{\mathbb{E}}_{Z_1 \sim \pi}\Big[\int_\Theta \mathbf{1}_{\{z_1 = z^* \wedge w(z^*, \theta) > 0\}} w(z_1, \theta)\pi_{z_1}\, \mathrm{d}\mu(\theta)\Big] \end{aligned}$$

$$= \mathop{\mathbb{E}}_{Z_1\sim\pi}\Big[\int_\Theta [w(z^*,\theta)\pi_{z^*}]_{\geq 0}\,\mathrm{d}\mu(\theta)\Big]$$
$$= \int_\Theta \int_{aux} \max_{\substack{z:\ a(z)=x \\ w(z,\theta)>0}} w(z_1,\theta)\pi_z\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(\theta).$$

Were the last inequality holds, since $z^* \in \arg\max_{a(z)=a(z_1)} w(\theta,z)\pi_z$. Applying Theorem 4.3 this is exactly the maximum RAD attainable. □

Theorem 4.9 directly establishes that Theorem 4.3 is universally tight and Theorem 4.2 is tight, since there exists at least one mechanism (GRR Theorem 4.4) for which Theorem 4.2 is tight for any auxiliary knowledge. We further validate that this is not an isolated case by empirically demonstrating tightness on additional mechanisms, such as DP-SGD (See Figure 4c).

The first major implication of tightness is that it enables the precise recalibration of DP mechanisms: one can optimize utility without sacrificing any unnecessary privacy margin, while retaining a clear and interpretable guarantee against reconstruction attacks. Second, beyond the theoretical contribution, our results provide a practical tool: a general attack algorithm that practitioners can directly use to evaluate the privacy risks of their systems or the tightness of their bounds. As a concrete demonstration, we apply this attack in the context of LDP auditing (see Section 6) and to assess empirical risk and tightness in private learning (see Section 7).

Example 4.10 (Optimal Attack on DP-SGD)**.** *Our analysis of DP-SGD is motivated by its central role in private learning: distributionally robust attacks were first introduced in this context [59], and DP-SGD remains the most widely used algorithm in practice [1]. In particular, we study the reconstruction setting considered by Hayes, Balle, and Mahloujifar [22], where the adversary attempts to reconstruct the target record z^* from a candidate set $\{z_1,\dots,z_m\}$ with uniform prior using access to the privatized gradients $\{\bar{g}_1,\dots,\bar{g}_T\}$ released during training, i.e., white-box setting.*

Given the output $\theta=(\theta_1,\dots,\theta_T)$, our optimal attack is determined by $\arg\max_{z:a(z)=x} w(\theta,z)$*, and its sign, i.e., whether $w(\theta,z)>0$ or not, for each candidate z and auxiliary knowledge x. Concretely, since the public dataset D_- is known, we can isolate the noisy contribution of the target's gradient at iteration t:*

$$g_t = \bar{g}_t - \sum_{z\in D_-} \mathrm{clip}_C(\nabla_\theta \ell(\theta_t,z)).$$

and simplify w maximization to

$$\arg\max_{z:a(z)=x} w(\theta,z) = \arg\max_{z:a(z)=x} \sum_t W(g_t, \mathrm{clip}_C(\nabla_\theta\ell(\theta_t,z))) \tag{43}$$

where $W(x,y)=\langle x,y\rangle - \frac{1}{m}\sum_z \langle x, \mathrm{clip}_C(\nabla_\theta\ell(\theta_t,z))\rangle$, since W preserves the sign and arg max *of w. We present the pseudo-code of the optimal attack in Algorithm 2.*

Indeed, given θ, z, under DP-SGD the privatized gradient at step t is

$$g_t \sim \mathcal{N}\Big(\mu_z,\, C^2\sigma^2 I\Big), \qquad \mu_z = \mathrm{clip}_C(\nabla_\theta\ell(\theta_t,z)),$$

where C is the clipping parameters and I the identity function of dimension d, corresponding to the dimension of the gradients. Hence the likelihood is

$$P_{\mathcal{M}}(g_t \mid z) \;=\; \underbrace{\frac{1}{(2\pi C^2\sigma^2)^{d/2}}}_{A} \exp\Big(\underbrace{-\frac{1}{2C^2\sigma^2}}_{B} \|g_t-\mu_z\|^2\Big),$$

where both A, B are independent from z. Consequently,

$$w(g,z) = \prod_t P_{\mathcal{M}}(g \mid z) - \prod_t P_{\mathcal{M}}(g_t) > 0 \Leftrightarrow \tag{44}$$

$$A^T\left(\prod_t \mathrm{e}^{B\langle g_t,\mu_z\rangle} - \prod_t \frac{1}{m}\sum_i \mathrm{e}^{B\langle g_t,\mu_{z_i}\rangle}\right) > 0 \Leftrightarrow \tag{45}$$

$$e^{B\sum_t \langle g_t, \mu_z\rangle} > \prod_t \frac{1}{m}\sum_i e^{B\langle g_t, \mu_{z_i}\rangle} \Leftrightarrow \tag{46}$$

$$B\sum_t \langle g_t, \mu_z\rangle > \sum_t \ln\left(\frac{1}{m}\sum_i e^{B\langle g_t, \mu_{z_i}\rangle}\right) \Leftrightarrow \tag{47}$$

$$B\sum_t \langle g_t, \mu_z\rangle > \sum_t \frac{1}{m}\sum_z \ln(e^{B\langle g_t, \mu_{z_i}\rangle}) \Leftrightarrow \tag{48}$$

$$B\sum_t \langle g_t, \mu_z\rangle > \sum_t \frac{B}{m}\sum_i \langle g_t, \mu_{z_i}\rangle \Leftrightarrow \tag{49}$$

$$\sum_t \langle g_t, \mu_z\rangle - \sum_t \frac{1}{m}\sum_z \langle g_t, \mu_z\rangle > 0 \Leftrightarrow \tag{50}$$

$$\sum_t W(g_t, z) > 0. \tag{51}$$

Where Equation (48) follows from the application of Jensen's inequality to the logarithm. Moreover, $\arg\max_z w(g,z) = \arg\max_z \ln(p_{\mathcal{M}}(g,z)) = \arg\max_z \sum_t p_{\mathcal{M}}(g_t, z)$*, where*

$$\ln p_{\mathcal{M}}(g_t \mid z_i) \propto -\tfrac{1}{2C^2\sigma^2}\,\|g_t - \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))\|^2.$$

Expanding the squared norm leads to

$$\|g_t\|^2 + \|\mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))\|^2 - 2\langle g_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))\rangle.$$

The term $\|g_t\|^2$ *is independent of* z_i*, and the term* $\|\mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))\|^2$ *is bounded by* C^2 *(often nearly constant across candidates). Therefore, maximizing the log-likelihood is equivalent to maximizing*

$$\langle g_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z_i))\rangle.$$

Consequently, our optimal attack can be simplified by using $W(g_t, z)$ *instead of* $w(g_t, z)$*.*

When $aux = \varnothing$*, our optimal attack coincides with the attack presented in [22]. Whereas they identified such an attack as the empirically best, we formally establish that this choice is indeed optimal. Moreover, we extend the optimal attack for any attacker that have target-specific auxiliary information. In particular, our optimal attack for attackers with* $aux \neq \varnothing$ *is empirically tested in Section 7, showing that previous bounds for ReRo indeed do not hold for attackers with target-specific auxiliary knowledge.*

Algorithm 2: Optimal Attack for DP-SGD

Input : $\theta = (\theta_1, \ldots, \theta_T)$, $a(z) = x$ *and* $g = (g_1, \ldots, g_T)$
Output : $\tilde{z}$
for $z\colon a(z) = x$ **do**
  compute $\sum_t W(\bar{g}_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z)))$;
Select $z^* = \arg\max_{z\colon a(z)=x} \sum_t W(\bar{g}_t, \mathrm{clip}_C(\nabla_\theta \ell(\theta_t, z)))\pi_z$;
if $W(\bar{g}_t, z^*) > 0$ **then**
  $\tilde{z} = z^*$;
else
  $\tilde{z} \longleftarrow U[\mathcal{Z}\backslash\{z\colon a(z) = x\}]$;

Summarizing, we have introduced a closed-form tight bound that holds independently of the adversary's auxiliary knowledge, thereby serving as a worst-case guarantee when the nature of *aux* is unknown (Theorem 4.2) and an auxiliary-dependent bound (Theorem 4.3) proved to be universally tight by constructing a general optimal attack strategy. We analyze our optimal attack as a risk-assessment and DP auditing tool in Sections 6 and 7.

Note that Theorem 4.3 imposes the hypothesis of disjoint success sets for every pair of records sharing the public auxiliary information. As previously discussed, this assumption is sensible for AIAs or categorical data reconstruction. However, reasonable intersections may indeed arise

in practice, especially in continuous data. For example, consider the task of reconstructing a person's salary using the ℓ_1 distance and allowing a sufficiently large error margin η. In this case, individuals sharing the same auxiliary information (e.g., the same profession) might have salaries within η of each other, thereby violating our assumption. Consequently, in the next section, we derive new bounds, covering this cases for the scenario in which the whole record is considered private and therefore, no auxiliary information about the target is available for the attacker.

5. η-RAD UPPER BOUNDS UNDER $aux = \varnothing$

Our bound in Theorem 4.3 is universally tight and accurate, but there are two aspects to improve. First, it relies on the assumption that reconstruction success sets do not intersect–a reasonable but imperfect hypothesis that may limit its applicability in some scenarios. Second, and more importantly for practical use, the bound is white-box: it requires full knowledge of the mechanism, and it does not have a closed form, which means that it may need to be numerically computed, especially for continuous data domains.

White-box bounds are useful for noise calibration; however, in DP auditing applications, we may only have query access to a mechanism (e.g., auditing an external software) without access to its internal protocol [19, 37]. Motivated by this, in this section, we focus on deriving closed-form, black-box bounds specifically for the case of $aux = \varnothing$–since it is the common assumption in prior DP auditing [4, 39] and data reconstruction studies [5, 22]. While these bounds hold for a weaker attacker model, they are directly applicable in black-box scenarios and remove the assumption of disjoint success subsets.

First, we present a general bound that applies both to any reconstruction setting as long as no target-specific auxiliary knowledge is available. For this purpose, we introduce $\kappa^-_{\pi,\ell}(\eta)$ as the infimum counterpart of $\kappa^+_{\pi,\ell}(\eta)$, formally defined as

$$\kappa^-_{\pi,\ell}(\eta) = \inf_{z_0 \in \mathcal{Z}} \Pr_{Z\sim\pi}\left[\ell(Z, z_0) \leq \eta\right], \tag{52}$$

which represents the success probability of an oblivious attacker attempting to reconstruct the most difficult target only using the prior π.

Theorem 5.1. *If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies f-DP, then any attack without target-specific auxiliary knowledge $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$ satisfies*

$$\eta\text{-RAD} \leq \max_{\alpha \in [\kappa^-_{\pi,\ell}(\eta), \kappa^+_{\pi,\ell}(\eta)]} 1 - f(\alpha) - \alpha.$$

If $\mathcal{Z}$ is discrete it also holds

$$\eta\text{-RAD} \leq (1 - \kappa_\pi) \max_{\alpha \in [0, \frac{\kappa^+_{\pi,\ell}(\eta)}{1-\kappa_\pi}]} 1 - f(\alpha) - \alpha.$$

Proof. Kifer et al. [31][p.23] that for any $S \subseteq \Theta$, for any f-DP mechanism, and $z_0, z_1 \in \mathcal{Z}$,

$$\Pr_{\mathcal{M}}(S \mid D_{z_1}) \leq 1 - f(\Pr_{\mathcal{M}}(S \mid D_{z_0}). \tag{53}$$

Moreover, since f is convex, applying Jensen's inequality:

$$f(\mathbb{E}_X[X]) \leq \mathbb{E}_X[f(X)] \Rightarrow -\mathbb{E}_X[f(X)] \leq f(\mathbb{E}_X[X]). \tag{54}$$

We denote $\mathbb{E}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\right] = 1 - \kappa_\pi = 1 - \Pr_{Z,Z'\sim\pi}[Z = Z']$, and apply these two properties obtaining,

$$\begin{aligned}
\eta\text{-RAD} &= \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_0,Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)] \\
&= \mathop{\mathbb{E}}_{Z_0\sim\pi}\Big[\Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr_{Z_1\sim\pi}[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\Big] \\
&= \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\Big[\mathbf{1}_{\{Z_0\neq Z_1\}}\big(\Pr[\mathcal{A}_{Z_1}(D_{Z_1}) \in S_\eta(Z_1)] - \Pr[\mathcal{A}_{Z_1}(D_{Z_0}) \in S_\eta(Z_1)]\big)\Big]
\end{aligned}$$

$$\leq \mathop{\mathbb{E}}_{Z_0,Z_1\sim\pi}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}(1-f(\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)])-\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)])\right]$$

$$=(1-\kappa_\pi)\left(1-\mathop{\mathbb{E}}_{Z_1,Z_0}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\tfrac{f(\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)])}{1-\kappa_\pi}\right]-\mathop{\mathbb{E}}_{Z_1,Z_0}\left[\mathbf{1}_{\{Z_0\neq Z_1\}}\tfrac{\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]}{1-\kappa_\pi}\right]\right).$$

In the continuous case $1-\kappa_\pi=1$, hence the expression of η-RAD's upper bound reduces to

$$\begin{aligned}
&1-\mathop{\mathbb{E}}_{Z_1,Z_0}[\mathbf{1}_{\{Z_0\neq Z_1\}}f\left(\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]\right)]-\mathop{\mathbb{E}}_{Z_1,Z_0}[\mathbf{1}_{\{Z_0\neq Z_1\}}\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]]\\
&\quad\leq 1-\mathop{\mathbb{E}}_{Z_1,Z_0}[f\left(\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]\right)]-\mathop{\mathbb{E}}_{Z_1,Z_0}[\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]]\\
&\quad\leq 1-f\left(\mathop{\mathbb{E}}_{Z_1,Z_0}[\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]]\right)-\mathop{\mathbb{E}}_{Z_1,Z_0}[\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]],
\end{aligned}$$

where last inequality follows from Equation (54). Therefore, it suffices to prove the interval where $\mathbb{E}_{Z_1,Z_0}[\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]$ lies,

$$\begin{aligned}
&\mathbb{E}_{Z_1,Z_0\sim\pi}\left[\Pr_{Z\sim\pi}[\mathcal{A}(D_{Z_0})\in S_\eta(Z)]\right]\\
&\quad=\int_{\mathcal{Z}}\int_{\mathcal{Z}}\Pr[\mathcal{A}(D_{z_0})\in S_\eta(z_1)]\pi_{z_0}\pi_{z_1}\,\mathrm{d}z_0\,\mathrm{d}z_1\\
&\quad=\int_{\mathcal{Z}}\int_{\mathcal{Z}}\int_{\mathcal{Z}}p_{\mathcal{A}}[z\mid D_{z_0}]\mathbf{1}_{\{\ell(z,z_1)\leq\eta\}}\pi_{z_0}\pi_{z_1}\,\mathrm{d}z_0\,\mathrm{d}z_1\,\mathrm{d}z\\
&\quad=\int_{\mathcal{Z}}\int_{\mathcal{Z}}p_{\mathcal{A}}[z\mid D_{z_0}]\left(\int_{\mathcal{Z}}\mathbf{1}_{\{\ell(z,z_1)\leq\eta\}}\pi_{z_1}\,\mathrm{d}z_1\right)\pi_{z_0}\,\mathrm{d}z_0\,\mathrm{d}z\\
&\quad\leq\kappa^+_{\pi,\ell}(\eta)\int_{\mathcal{Z}}\int_{\mathcal{Z}}p_{\mathcal{A}}[z\mid D_{z_0}]\pi_{z_0}\,\mathrm{d}z_0\,\mathrm{d}z=\kappa^+_{\pi,\ell}(\eta).
\end{aligned}$$

and analogous for $\kappa^-_{\pi,\ell}(\eta)$ since any attack output $z\in\mathcal{Z}$ and hence it follows the definition. Note that, last inequality assumes no auxiliary knowledge is available there therefore $p_{\mathcal{A}}[z\mid D_{z_0},a(z_1)]=p_{\mathcal{A}}[z\mid D_{z_0}]$, hence it factors out of the integral respect to z_1.

For the discrete case, $(1-\kappa_\pi)\neq 1$, but $\sum_{z_1}\sum_{z_0\neq z_1}\frac{\pi_0\pi_1}{(1-\kappa_\pi)}=1$, hence applying Jensen's inequality

$$\begin{aligned}
&(1-\kappa_\pi)\left(1-\mathop{\mathbb{E}}_{Z_1,Z_0}[\mathbf{1}_{\{Z_0\neq Z_1\}}\frac{f\left(\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]\right)}{(1-\kappa_\pi)}]-\mathop{\mathbb{E}}_{Z_1,Z_0}[\mathbf{1}_{\{Z_0\neq Z_1\}}\frac{\Pr[\mathcal{A}_{Z_1}(D_{Z_0})\in S_\eta(Z_1)]}{(1-\kappa_\pi)}]\right)\\
&\leq(1-\kappa_\pi)\left(1-f(\sum_{z_1}\sum_{z_0\neq z_1}\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]\tfrac{\pi_0\pi_1}{(1-\kappa_\pi)})-\sum_{z_1}\sum_{z_0\neq z_1}\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]\tfrac{\pi_0\pi_1}{(1-\kappa_\pi)}.\right)
\end{aligned}$$

Therefore, the proof follows from the following upper-bound:

$$\sum_{z_1}\sum_{z_0\neq z_1}\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]\tfrac{\pi_0\pi_1}{(1-\kappa_\pi)}\leq\frac{1}{(1-\kappa_\pi)}\mathop{\mathbb{E}}_{Z_0,Z_1}[\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]=\frac{\kappa^+}{(1-\kappa_\pi)}.\quad\square$$

This result serves as an upper-bound approximation of RAD, when $S_\eta(z)\cap\mathcal{S}_\eta(z')\neq\varnothing$ (hence we can not apply Theorem 4.3) and $aux=\varnothing$. Moreover, as a consequence of the previous result, we can obtain its black-box counterpart, i.e., a general result for any (ε,δ)-DP mechanism even if the mechanism itself is unknown:

Proposition 5.2. *If a mechanism $\mathcal{M}\colon\mathcal{Z}^n\to\mathcal{D}(\Theta)$ satisfies (ε,δ)-DP, then for any attack $A\colon\Theta\to\mathcal{D}(\mathcal{Z})$,*

$$\eta\text{-}RAD\leq\min\{\kappa^+_{\pi,\eta}(\mathrm{e}^\varepsilon-1)+\delta,\frac{1-\delta-\kappa^-(1+\mathrm{e}^\varepsilon)}{\mathrm{e}^\varepsilon},\frac{e^\varepsilon-1+2\delta}{e^\varepsilon+1}(1-\kappa_\pi)\}.$$

Proof. Follows from combining previous theorem with [11] result that any (ε,δ)-DP mechanism is f-DP with, $f(\alpha)=\max\{1-\delta-\mathrm{e}^\varepsilon\alpha,\frac{1-\delta-\alpha}{\mathrm{e}^\varepsilon}\}$, and analyze the different cases until we arrive

to the bound. Formally, every (ε, δ)-DP mechanism verifies the that f-DP, with f

$$f(\alpha) = \max\{\underbrace{1 - \delta - \mathrm{e}^{\varepsilon}\alpha}_{f_1(\alpha)}, \underbrace{\frac{1 - \delta - \alpha}{\mathrm{e}^{\varepsilon}}}_{f_2(\alpha)}\}. \tag{55}$$

On the other side, applying Theorem 5.1 we have

$$\eta\text{-RAD} \leq \max_{\alpha \in [\kappa^-, \kappa^+]} (1 - f(\alpha) - \alpha)\,. \tag{56}$$

Combining both equations we obtain,

$$\begin{aligned}
\eta\text{-Rad} &\leq \max_{\alpha \in [\kappa^-, \kappa^+]} (1 - f(\alpha) - \alpha) \\
&= \max_{\alpha \in [\kappa^-, \kappa^+]} 1 - \max\{f_1(\alpha), f_2(\alpha)\} - \alpha \\
&= \max_{\alpha \in [\kappa^-, \kappa^+]} (1 - \max\{f_1(\alpha) + \alpha, f_2(\alpha) + \alpha\}) \\
&= \max_{\alpha \in [\kappa^-, \kappa^+]} (\min\{1 - f_1(\alpha) - \alpha, 1 - f_2(\alpha) - \alpha\}) \\
&\leq \min\{\max_{\alpha \in [\kappa^-, \kappa^+]} 1 - f_1(\alpha) - \alpha, \max_{\alpha \in [\kappa^-, \kappa^+]} 1 - f_2(\alpha) - \alpha\}
\end{aligned}$$

Therefore, we analyze both maximums,

First, for f_1 we have:

$$1 - f_1(\alpha) - \alpha = \delta + \mathrm{e}^{\varepsilon}\alpha - \alpha \tag{57}$$

$$= \alpha(\mathrm{e}^{\varepsilon} - 1) + \delta \leq \kappa^+(\mathrm{e}^{\varepsilon} - 1) + \delta \tag{58}$$

Second, for f_2 we obtain:

$$1 - f_2(\alpha) - \alpha = 1 - \frac{1 - \delta - \alpha}{\mathrm{e}^{\varepsilon}} - \alpha \tag{59}$$

$$= 1 - \frac{1 - \delta}{\mathrm{e}^{\varepsilon}} + \alpha(\mathrm{e}^{-\varepsilon} - 1) \leq 1 - \kappa^-(1 - \mathrm{e}^{-\varepsilon}) - \frac{1 - \delta}{\mathrm{e}^{\varepsilon}} = \frac{1 - \delta - \kappa^-(1 + \mathrm{e}^{\varepsilon})}{\mathrm{e}^{\varepsilon}}. \tag{60}$$

Combined with the general bound Theorem 4.2 it follows the result.

□

Next, we focus on perfect reconstruction, i.e. $\eta = 0$, in categorical data. This case is particularly relevant since many sensitive attributes, such as diseases, political opinions, or religious beliefs, are categorical in nature. This case does not trivially support partial reconstruction, e.g. [17, 18]. For such settings, we can derive more precise bounds. To do so, we first introduce the following auxiliary lemma:

Lemma 5.3. *Given $|\mathcal{Z}| = m$ and $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$*

$$\Gamma \coloneqq \int_{\theta} \arg\max_{z} w(z, \theta)\, \mathrm{d}\mu(\theta) \leq \frac{(m-1)(\mathrm{e}^{\varepsilon} - 1 + \delta m)}{\mathrm{e}^{\varepsilon} + m - 1}. \tag{61}$$

Proof. First, since $|\mathcal{Z}| = m$, we obtain

$$\int_{\Theta} \arg\max_{z \in \mathcal{Z}} p_{\mathcal{M}}(\theta \mid z)\, \mathrm{d}\mu(\theta) = \sum_{z \in \mathcal{Z}} \int_{\Theta(z)} p_{\mathcal{M}}(\theta \mid z)\, \mathrm{d}\mu(\theta) = \sum_{z \in \mathcal{Z}} \Pr_{\mathcal{M}}(\Theta_z \mid z), \tag{62}$$

where Θ_z are recursively defined as in Equation (18). By definition $\Theta_z \cap \Theta_{z'} = \varnothing$. Besides, since for all θ it exits at least one $z_\theta \in \arg\max_z p_{\mathcal{M}}(\theta \mid z)$, $\cup \Theta_z = \Theta$. Hence, $\{\Theta_z\}_{z \in \mathcal{Z}}$ determines a partition in Θ. Therefore, by the law of total probability, for each z_0 we have

$$\sum_{z \in \mathcal{Z}} \Pr_{\mathcal{M}}(\Theta_z \mid z_0) = \sum_{z} \int_{\Theta_z} p_{\mathcal{M}}(\theta \mid z_0)\, \mathrm{d}\mu(\theta) = \int_{\Theta} p_{\mathcal{M}}(\theta \mid z_0)\, \mathrm{d}\mu(\theta) = 1. \tag{63}$$

On the other hand, since $\mathcal{M}$ is (ε, δ)-DP, for every $z_1, z_0 \in \mathcal{Z}$,

$$\Pr_{\mathcal{M}}(\Theta_1 \mid z_0) \geq \mathrm{e}^{-\varepsilon}(\Pr_{\mathcal{M}}(\Theta_1 \mid z_1) - \delta). \tag{64}$$

substituting Equation (64) in Equation (63) we obtain, for all $i, j \in [m]$,

$$\Pr_{\mathcal{M}}(\Theta_i \mid z_i) + \mathrm{e}^{-\varepsilon} \sum_{i \neq j} \Pr_{\mathcal{M}}(\Theta_j \mid z_j) \leq 1 + \delta \mathrm{e}^{-\varepsilon}(m-1) \tag{65}$$

Summing the above inequality over all $i \in [m]$,

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) + (m-1)\mathrm{e}^{-\varepsilon} \sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) \leq m(1 + \delta \mathrm{e}^{-\varepsilon}(m-1)) \Leftrightarrow \tag{66}$$

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) \leq \frac{m(1 + \delta \mathrm{e}^{-\varepsilon}(m-1))}{1 + (m-1)\mathrm{e}^{-\varepsilon}} = \frac{m\mathrm{e}^{\varepsilon} + \delta m(m-1)}{\mathrm{e}^{\varepsilon} + (m-1)}. \tag{67}$$

Hence,

$$\Gamma = \int_{\theta} \arg\max_{z} w(z, \theta) \, \mathrm{d}\mu(\theta) \tag{68}$$

$$= \int_{\theta} \left(\arg\max_{z} p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta) \right) \mathrm{d}\mu(\theta) \tag{69}$$

$$= \int_{\theta} \arg\max_{z} p_{\mathcal{M}}(\theta \mid z) \, \mathrm{d}\mu(\theta) - 1 \tag{70}$$

$$\leq \frac{m\mathrm{e}^{\varepsilon} + \delta m(m-1)}{\mathrm{e}^{\varepsilon} + m - 1} - 1 = \frac{(m-1)(\mathrm{e}^{\varepsilon} - 1 + \delta m)}{\mathrm{e}^{\varepsilon} + m - 1}. \tag{71}$$

□

Applying this lemma we obtain the following RAD bound:

Theorem 5.4 (0-RAD under (ε, δ)-DP)**.** *Given π distribution over $|\mathcal{Z}| = m$, such that $\pi_1(1-\pi_1) \geq \pi_2(1-\pi_2) \geq \cdots \geq \pi_m(1-\pi_m)$ and $\mathcal{M} \colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A \colon \Theta \to \mathcal{D}(\mathcal{Z})$*

$$\text{0-RAD} \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} K_{\pi} + R \max_{i \notin K} \pi_i$$

where the biggest $K \in [m]$ such that $R = (m-1)\frac{e^{\varepsilon}-1+m\delta}{e^{\varepsilon}+m-1} - (K - \sum_{i=1}^{K} \pi_i)\frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1} \geq 0$ and $K_{\pi} = \sum_{i}^{K}(1-\pi_i)\pi_i$.

Proof. Since $|\mathcal{Z}| = m$ and $aux = \varnothing$, Theorem 4.3 gets reduced to *Equation* (18), hence

$$\text{0-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) \right) \pi_i \equiv \sum_{i=1}^{m} \gamma_i \, \pi_i. \tag{72}$$

For one side, we obtain that for all $i \in [m]$,

$$\gamma_i = \Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) = \int_{\Theta_i} p_{\mathcal{M}}(\theta \mid z_i) - \sum_{j \in [m]} p_{\mathcal{M}}(\theta \mid z_j)\pi_j \, \mathrm{d}\mu(\theta) \tag{73}$$

$$= \int_{\Theta_i} \sum_{j \in [m]} \left(p_{\mathcal{M}}(\theta \mid z_i) - p_{\mathcal{M}}(\theta \mid z_j) \right) \pi_j \, \mathrm{d}\mu(\theta) \tag{74}$$

$$= \sum_{j \neq i} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i \mid z_j) \right) \pi_j \tag{75}$$

$$\leq \mathrm{TV}(\mathcal{M}) \sum_{j \neq i} \pi_j \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(1 - \pi_i). \tag{76}$$

If we simply apply this bound we recover Theorem 4.2 result:

$$\text{0-RAD} \leq \sum_{i=1}^{m} \gamma_i \, \pi_i \leq \sum_{i=1}^{m} \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(1 - \pi_i)\pi_i = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(1 - \kappa_{\pi}).$$

However, due to Theorem 5.3, we know that this bound is loose, since in this case,

$$\Gamma = \sum_{i=1}^{m} \gamma_i = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1}(m-1) \geq \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1}(m-1) = \Gamma_{\max}, \tag{77}$$

contradicting Theorem 5.3; therefore, it is impossible to achieve the local inequality $\gamma_i \leq \mathrm{TV}(\mathcal{M})(1-\pi_i)$ simultaneously for all $i \in [m]$. In most cases, cases, we can apply the local bound to a reduced set of indexes k, and the reminders must adjust so that the total sum $\sum_i \gamma_i = \Gamma$. Formally, at most, we can sum k summands such that,

$$\sum_{r=1}^{k} \gamma_{i_r} \leq \frac{e^\varepsilon - 1 + m\delta}{e^\varepsilon + m - 1}(m-1) \Leftrightarrow \tag{78}$$

$$\sum_{r=1}^{k}(1-\pi_{i_r}) \leq (m-1)\frac{(\mathrm{e}^\varepsilon - 1 + m\delta)((\mathrm{e}^\varepsilon + 1)}{(\mathrm{e}^\varepsilon - 1 + 2\delta)((\mathrm{e}^\varepsilon - 1 + 2\delta))} \tag{79}$$

Hence, without lost of generality we order the indices so that

$$\pi_1(1-\pi_1) \geq \pi_2(1-\pi_2) \geq \cdots \geq \pi_m(1-\pi_m).$$

obtaining,

$$0\text{-RAD} \leq \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}\sum_{i=1}^{k_\pi} \pi_i(1-\pi_i) + R \max_{r>k_\pi} \pi_r \tag{80}$$

with k_π the maximum index verifying:

$$\sum_{i=1}^{k_\pi}(1-\pi_i) \leq (m-1)\frac{(\mathrm{e}^\varepsilon - 1 + m\delta)((\mathrm{e}^\varepsilon + 1)}{(\mathrm{e}^\varepsilon - 1 + 2\delta)((\mathrm{e}^\varepsilon - 1 + 2\delta))}.$$

and R the reminder, i.e, K the biggest index such that

$$R = (m-1)\frac{e^\varepsilon - 1 + m\delta}{e^\varepsilon + m - 1} - (K - \sum_{i=1}^{K}\pi_i)\frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1} \geq 0. \qquad \square$$

Note that in the extreme case where $\pi_1 = \pi_2 = \frac{1}{2}$ and $\pi_i = 0$ for all $i \neq 1, 2$, we recover exactly the same result as in Theorem 4.2. This formulation enables the assessment of intermediate configurations of π. Notably, when $\pi = U[m]$ yields a marked improvement:

Corollary 5.5 (Black-box Uniform Prior)**.** *Let $\pi = \mathrm{U}[m]$ the uniform distribution over $\mathcal{Z}$. If a mechanism $\mathcal{M}$ satisfies (ε, δ)-DP, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$ it guarantees*

$$0\text{-RAD} \leq \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}.$$

Proof. For every $K \in [m]$, $K_\pi = \sum_{i=1}^{K}(1-\pi_i)\pi_i = K\frac{m-1}{m^2}$ and $(K - \sum_{i=1}^{K}\pi_i) = K\frac{m-1}{m}$, therefore, denoting $A = \frac{\mathrm{e}^\varepsilon - 1 + 2\delta}{\mathrm{e}^\varepsilon + 1}$ and applying Theorem 5.4 we get:

$$0\text{-RAD} \leq AK\frac{m-1}{m^2} + \frac{1}{m}(\Gamma - K\frac{m-1}{m}A) = \frac{1}{m}\Gamma = \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}. \tag{81}$$

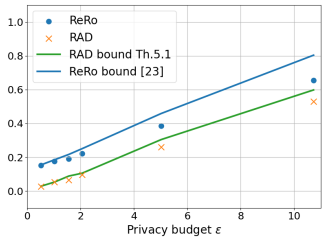
$\square$

In this section, we presented closed-form black-box bounds that rely only on the privacy parameters, making them especially relevant in settings where the mechanism is not disclosed, such as black-box auditing scenarios [37, 43]. Moreover, Theorems 5.4 and 5.5 provide more accurate estimation than Theorems 4.2 and 5.2 for application in categorical data.

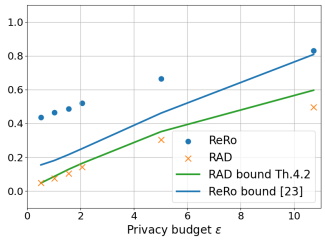
6. RAD for DP Auditing

DP auditing is crucial for assessing the tightness of DP mechanisms, establishing the practical impact of the mechanism parameters, and detecting implementation flaws in deployed DP mechanisms [7, 25]. While previous DP auditing tools focus on solving specifically one of the aforementioned aspects, we propose a general-purpose DP auditing framework: RAD-based DP auditing.

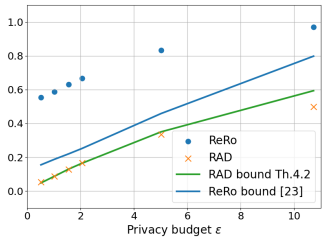
RAD provides a unifying framework for analyzing adversarial risk under arbitrary threat models. Moreover, our bounds establish a tight and explicit connection between RAD and the standard privacy parameters. Taken together, these results yield a simple and principled



(A) DRA, no target-specific auxiliary knowledge $aux = \varnothing$.



(B) DRA, auxiliary knowledge : $a(z)$ = image label.



(C) MIA, full knowledge about the target record: $a(z) = z$.

FIGURE 4. RAD vs ReRo results for optimal attacks against DP-SGD on MNIST. Lines show theoretical bounds and markers of empirical risk as estimated by RAD/ReRo. Empirical results exceed the bounds as estimated by ReRo, RAD bounds hold.

approach to general-purpose DP auditing. Precision and tightness are especially critical in this context, since loose estimates may underestimate privacy risks or fail to detect bugs and implementation flaws.

The core idea of RAD-based auditing is straightforward: given a measured RAD value, we invert our theoretical bound to estimate an empirical privacy budget. This empirical $\tilde{\varepsilon}$ reflects the observed privacy loss in practice, complementing theoretical worst-case values and providing a more realistic perspective on real-world risk. Formally, given a bound η-RAD$\leq B(\varepsilon, \delta)$, we compute RAD empirically obtaining γ, and estimate $\tilde{\varepsilon} \geq B^{-1}(\gamma, \delta)$.

The bound we employ depends on the specific setting. For instance, in a black-box scenario for categorical data, in which we assume $\pi = U[m]$, the best bound is Theorem 5.5. Therefore, the DP auditing framework consists of running an attack, measuring its empirical RAD $\widetilde{\gamma}$, and deriving $\widetilde{\varepsilon}$ as follows:

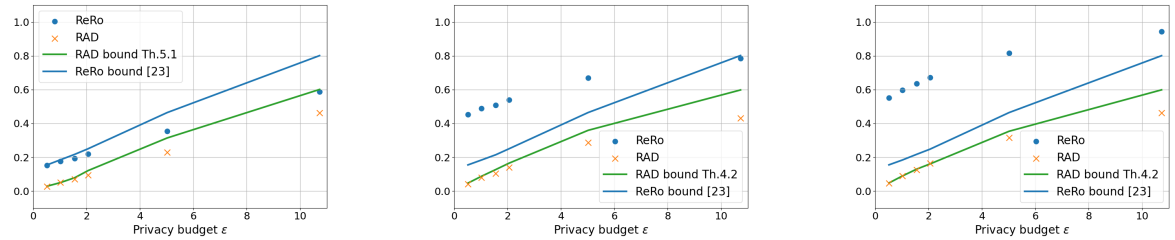
$$\widetilde{\varepsilon} = \begin{cases} \ln\left(\frac{\widetilde{\gamma}\frac{m}{m-1}+1}{1-\widetilde{\gamma}\frac{m}{m-1}}\right) & \text{if the term can be evaluated,} \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{82}$$

However, for white-box auditing, we can use our improved bound from Theorem 4.3.

Despite the importance of LDP mechanisms [16], only one major work has so far focused on LDP auditing: LDP AUDITOR [4]. Applying our RAD-based DP auditing to LDP we address key limitations of prior work. In contrast to LDP AUDITOR, which focuses exclusively on perfect reconstruction without target-specific auxiliary knowledge–excluding important use-cases such as AIAs–we allow auditing under broader threat models by leveraging optimal attacks (see Algorithm 1). Moreover, our approach is not constrained by internal parameter choices that bound the maximum privacy loss estimate (as in LDP AUDITOR) [4], thus providing tighter and more accurate guarantees. We investigate and empirically show the improvement in accuracy of our auditing approach in Section 7 (cf. Figure 6 for results), where we audit three main LDP mechanisms–GRR, SS and OUE–showing improved accuracy for all of them.

7. EXPERIMENTS

In this section, we empirically examine the limitations of ReRo described in Section 3, focusing on how existing bounds fail to account for realistic attackers with target-specific auxiliary information. Moreover, we validate our theoretical bounds and our RAD-based DP auditing framework in real-world databases and DP mechanisms. Our experiments show that RAD accurately distinguishes privacy leakage from imputation, with tight bounds in practice, making it a reliable tool for interpretable noise calibration. RAD also enables auditing of LDP mechanisms, improving both scope and accuracy over the state-of-the-art [4].



(A) DRA, no target-specific auxiliary knowledge $aux = \varnothing$.

(B) DRA, auxiliary knowledge : $a(z)$ = image label.

(C) MIA, full knowledge about the target record: $a(z) = z$.

FIGURE 5. RAD vs ReRo results for optimal attacks against DP-SGD on fashion. Lines show theoretical bounds and markers of empirical risk as estimated by RAD/ReRo. Empirical results exceed the bounds as estimated by ReRo, RAD bounds hold.

7.1. **Database Description.** We evaluate both private learning and LDP scenarios, using tailored datasets for each setting. The database selection is guided by their relevance in prior work and availability.

For the analysis of optimal attacks against DP-SGD, we use the same dataset as in the evaluation of ReRo [22] for consistency–namely, MNIST [33], a benchmark of 70,000 grayscale images of handwritten digits (0–9). We further replicate our results on Fashion-MNIST [56] (referred to as Fashion), which, likewise, contains 70,000 grayscale images, but of clothing items.

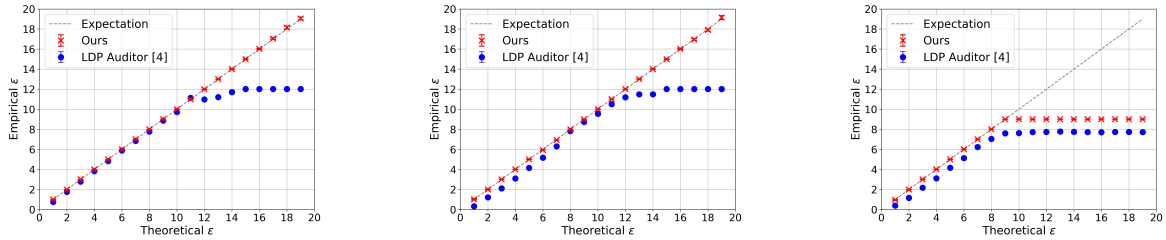
To evaluate the imputation attack [28], we use the Census and Texas-100X datasets in consistency with the original paper. The Census dataset [28] contains 1,676,013 records with 14 attributes, where race is treated as the sensitive attribute with eight categories. The Texas-100X dataset [28] comprises 925,128 patient records from 441 hospitals, including demographic and medical attributes, with a binary ethnicity attribute designated to be sensitive.

We evaluate our LDP auditing framework on location-reconstruction attacks using two real-world mobility datasets: the Porto dataset [47] and the Geolife dataset [60]. Both datasets are widely used in privacy and mobility research (e.g., [35, 49, 57]) and are publicly available. Each dataset consists of GPS coordinates which we map to the OpenStreetMap (OSM) graph format [48] like prior work. The Porto dataset contains a total of $83,409,386$ location reports that we map to the OSM roadgraph at Porto's city center (41.1475, -8.5870) with a 2.7 km radius, capturing the urban core of Porto. This radius leads to a universe size $|\mathcal{Z}| = 3,052$. The Geolife dataset contains a total of $24,876,978$ locations that we mapped to an OSM graph centered near Tiananmen Square (39.9130, 116.3703) with a 5 km radius covering major central districts, leading to a universe of size $|\mathcal{Z}| = 5,356$.

7.2. **Experiment Design.** We explore attacks on DP-SGD and LDP auditing covering different auxiliary information settings—to validate our different bounds—and empirically apply our optimal attacks.

We expose the risk overestimation of ReRo due to imputation and how RAD overcomes this issue with the pure imputation attack [28], which uses a public dataset D_- to train a separate attack classifier A_I that, given the public attributes of a target, returns as label a prediction for the sensitive one. The adversary is given only the target public attribute $a(z)$ and outputs the prediction $\widetilde{s_z} = \arg\max_{s_i \in \Theta} \Pr_{\mathcal{I}}[s_i \mid a(z)]$, where the conditional distribution $\Pr[s_i \mid a(z)]$ is estimated by A_I, once the imputation model has been trained on D_-. This attack does not use any information from the target model $\mathcal{M}(D)$; therefore, adversarial success cannot be privacy leakage resulting from a user's participation in the training dataset of $\mathcal{M}(D)$. Following the original paper [28], we tested in both the Census and Texas datasets. We set $|D_-| = 49,000$ and a universe $\mathcal{Z}$ of $m = 1000$, randomly selected from the remaining data records consistent with [28]. We define the attack to be successful, $\ell(z, z') = 0$, if $a(z) = a(z')$, as typical for AIAs.

We evaluate our optimal attack strategy (see Algorithm 2) on the MNIST and Fashion image datasets in three experimental settings: $aux = z$ (a MIA), $aux = \varnothing$ (a DRA, replicating the



(A) (GRR) (B) (SS) (C) (OUE)

FIGURE 6. LDP Audit results from RAD-based auditing and LDP AUDITOR [4] on Porto dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

setting from [22]), and $aux = a(z)$ (a DRA, where the adversary additionally knows the target image's label, i.e., which object the image represents). We declare an attack successful when $A(\theta, a(z)) = z$, that is, $\eta = 0$.

We set $|D_-| = 999$ (and so the training set size is $|D_- \cup \{z\}| = 1,000$) and train with full-batch DP-SGD for $T = 100$ steps. We set $C = 0.1$ and $\delta = 10^{-5}$ and adjust the noise scale σ for a given target ε. We set the uniform prior with size $|\mathcal{Z}| = 8$ (disjoint from D_-), meaning that $\kappa^+_{\pi,0} = \kappa_\pi = 0.125$. Hence, we exactly replicate the original ReRo study [22] parameters.

Finally, we evaluate RAD in LDP and we compare our auditing framework with the state-of-the-art tool LDP AUDITOR [4] for three relevant LDP mechanisms: GRR, OUE and SS [4, 21]. To obtain the results for LDP AUDITOR, we used the code from Arcolezi and Gambs's public GitHub repository [3]. LDP AUDITOR estimates the empirical privacy budget in 10^6 runs.

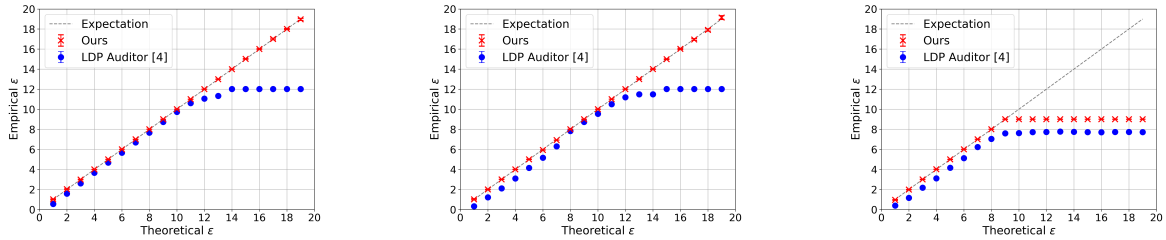
We evaluate RAD based on our optimal attack (See Algorithm 1) under a uniform prior and without auxiliary knowledge, allowing comparison with LDP AUDITOR. We then test our own LDP auditing framework: based on the obtained RAD value γ, we evaluate $B^{-1}(\gamma)$ for B following Theorem 4.3 and obtain an estimate of the empirical privacy budget. The precise $B(\varepsilon)$ for GRR, OUE and SS are shown in Theorems 4.4 to 4.6 respectively. Since B^{-1} is not explicit for OUE, we approximate it numerically using the bisection method, which converges in $\mathcal{O}(\log(\tau^{-1}))$ iterations, where τ denotes the tolerance level [50]. We set $\tau = 10^{-6}$. Consistently with [4], we repeat the epsilon estimation five times and report the mean and standard deviation.

All of our experiments are based on empirical estimates of ReRo and RAD. To estimate the ReRo, we follow [22]: Repeat the attack $A(\mathcal{M}(D_- \cup \{z\}), a(z))$, J times for each target record $z \in \mathcal{Z}$ and average the results (since π is assumed uniform). We follow the same procedure to estimate the correction term in RAD, i.e, for each pair of target-challenger $z_1, z_0 \in \mathcal{Z}$ we compute $A(\mathcal{M}(D_{z_0}, a(z_1))$ J times and average the results.

For MNIST and Fashion, we set $J = 1,000$ (as in [22]). Note that in the LDP cases $D_- = \varnothing$, and we set $J = 10^6/m$ ensuring the total number of runs matches those 10^6 repetitions of LDP AUDITOR. Finally, for the imputation attack, we do not require a target model as it is target model-independent and set $J = 1$. We repeat the imputation attack with five different seeds and report the averaged ReRo and RAD scores.

We use Python and the ML library Tensorflow [53] to evaluate the attacks. We use a minimal implementation of kindly provided by Hayes, Balle, and Mahloujifar for DP-SGD ReRo computation, and we extend their experiments to incorporate RAD and target-specific auxiliary knowledge. For the imputation attack [28], we adapt the code implementation from the authors' public repository [27].

7.3. **Results.** In this section, we present the results of our optimal attacks against DP-SGD and LDP mechanism, including both RAD and ReRo empirical risk estimates together with their corresponding theoretical bounds. Note that for both ReRo and RAD, the y-axis represents the risk measure, where values close to one indicate high risk, while values near zero indicate low risk.



(A) (GRR) (B) (SS) (C) (OUE)

FIGURE 7. LDP Audit results from RAD-based auditing and LDP AUDITOR [4] on Geolife dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

Dataset	ReRo	RAD
Census	0.81	0
Texas	0.73	0

TABLE 1. ReRo Vs. RAD risk estimation for imputation attack.

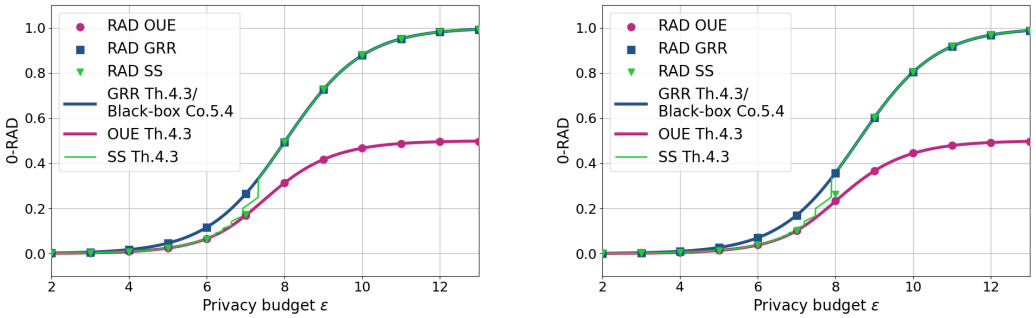
7.3.1. *RAD covers, but ReRo breaks for auxiliary knowledge.* Figure 4 shows the results of ReRo and RAD risk estimation for our optimal attacks against DP-SGD on the MNIST dataset. Analogous results for the Fashion dataset are provided in Figure 5 of the paper. We also include the corresponding theoretical bounds for ReRo and RAD for comparison. As expected, the existing bounds for ReRo [22] correctly upper-limit the empirically observed ReRo risk when the adversary has no prior knowledge of the victim record ($aux = \varnothing$, Figure 4a). However, when the adversary has prior knowledge of the victim record (Figures 4b and 4c), ReRo reveals higher disclosure than predicted by its theoretical bounds. In contrast, our RAD bounds consistently upper-limit the empirically estimated RAD risks across all tested attacks.

This supports our expectation that the ReRo bound only holds under the assumption that the adversary has no auxiliary knowledge about the victim ($aux = \varnothing$), but fails to correctly estimate privacy risks when target-specific auxiliary knowledge exists.

We can also observe that our bounds for RAD overcome this estimation error: they hold for any auxiliary knowledge and are nearly tight. In particular, Figures 4b and 4c show that the tightness of our worst-case bound Theorem 4.2 is not an isolated feature of GRR, but a reliable property that also applies to other widely used mechanisms, such as DP-SGD. Finally, Figure 4a shows that our closed-form bound Theorem 5.1 offers reasonable upper-bound also when Theorem 4.3 needs to be numerically approximated (as is the case, for instance, with DP-SGD).

7.3.2. *Leakage vs. Imputation.* Table 1 compares the risk estimates of RAD and ReRo for the imputation attack. This attack is not based on any information leakage from the mechanism and ignores any output in the process. RAD in this case does estimate the privacy risk to be 0, whereas ReRo reports notably higher values (0.81 for Census and 0.73 for Texas). This underlines how RAD is the more reliable measure of actual privacy risks: RAD shows the absence of leakage when the attack's success relies solely on imputation, whereas ReRo suggests serious disclosures (or: attack potential), effectively overestimating the privacy risk.

The tendency of ReRo to overestimate risk is not limited to this case. Going back to the results of our optimal attacks on DP-SGD (Figure 4) we observe that ReRo overestimates the leakage in all other investigated cases as well. This overestimation becomes more pronounced as additional auxiliary information is taken into account. Membership inference ($a(z) = z$) represents the clearest example in our study of ReRo estimating exceedingly high risk values of over 0.6, even for budgets $\varepsilon \leq 4$, which are typically regarded in the literature as providing high levels of privacy [34]. These observations meet our expectations well, as the inability to discount



(A) Porto (B) Geolife

FIGURE 8. RAD results for LDP mechanisms. Lines show theoretical bounds and markers empirical RAD.

for auxiliary information is one of the reasons for ReRo to overestimate risk. Consequently, the more auxiliary information the attacker possesses, the greater the degree of overestimation we observe for ReRo.

7.3.3. *RAD in LDP mechanisms.* **??** shows the results of our optimal attack analysis against LDP mechanisms, GRR, OUE and SS, on the Porto and Geolife datasets, including RAD empirical risk estimates together with their corresponding theoretical bounds. On the x-axis we see the privacy budget ε and the y-axes the exact RAD estimated risk for such ε selection. We see that our bounds (corresponding to Theorem 4.3) are perfectly tight and capture even the most subtle differences between mechanisms. Notably, RAD estimates for GRR perfectly align with our perfect-reconstruction black-box bound Theorem 5.4, confirming its tightness. Moreover, they provide a tangible argument for the importance of attack-based noise calibration: for the same values of ε, OUE offers significantly better protection against reconstruction attacks than GRR and SS. Therefore, ε does not capture the whole picture, and RAD is crucial for understanding the actual privacy implications of a mechanism on users.

7.3.4. *Auditing Local DP with RAD.* Figure 6 shows the results from our LDP Auditing experiments using the Porto dataset (experiments on the Geolife dataset yield similar results, which we show in Figure 7). They compare the accuracy of predicting the actual ε using our RAD-based auditing versus LDP AUDITOR. The closer the empirical ε is to the theoretical value (diagonal line), the more accurate the auditing tool. Additionally, smaller standard deviations indicate greater stability of the method.

For all tested mechanisms, our auditing approach improves over LDP AUDITOR for all ε values. In particular, we see that the highest ε LDP AUDITOR manages to estimate for both GRR and SS are capped around $\widetilde{\varepsilon} \approx 12.25$, hence preventing auditing of deployments with higher values. This limitation was already acknowledged by the authors of LDP AUDITOR, as it stems from the intrinsic shortcomings of the Clopper–Pearson method underlying their approach [4] In contrast, the tightness of our attribute advantage bound enables our auditing approach to accurately estimate empirical privacy budgets for the whole range, without such limitation. Notably, for GRR and SS, our DP auditing yields near perfect estimates for all epsilon values. For the OUE mechanism, our approach also outperforms LDP AUDITOR, however, the estimation accuracy declines at $\varepsilon \leq 9$. Note that this is an inherent limitation of OUE auditing as already mentioned in [4]: as we prove in Theorem 4.5, 0-RAD converges to $\frac{m-1}{2m}$ when ε tends to infinity. Overall, these results support that the universal tightness of our theoretical bound Theorem 4.3 enables precise and reliable auditing based on reconstruction attacks.

8. CONCLUSION

In this paper, we investigate the reconstruction risk that users incur when their data are processed by DP mechanisms. Our results reveal that the current state-of-the-art risk metric, ReRo [5], drastically overestimates the actual leakage of DP mechanisms when target-specific

public knowledge exists. Crucially, we show that under real attacks, existing ReRo bounds are violated.

To address these limitations, we first introduce η-RAD, a novel metric consistent with attribute and membership advantage, that accurately captures the privacy risk imposed by any specific mechanism. More importantly, we advance the understanding and practical interpretation of DP guarantees by proving tight bounds that connect DP mechanisms with their risk, using RAD. Offering new insights and clarity beyond existing analyses, we establish (i) universally tight bounds when the attacker's knowledge is specified, along with optimal strategies achieving them, (ii) closed-form bounds that remain valid regardless of auxiliary knowledge, and (iii) black-box upper bounds for settings with completely secret records. Our theoretical and empirical evaluation–across both LDP and private learning settings–demonstrates not only the robustness of RAD as a risk measure, but also the significant impact of our bounds on improving DP noise calibration (proving better utility) and auditing in DP (broadening the scope and improving accuracy).

Overall, our work demonstrates that privacy risk depends on the mechanism's structure, not just its nominal privacy parameters, and provides both fundamental insight and practical tools for privacy risk assessment and calibration – enabling notable utility gains without increasing the effective privacy risk.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. "Deep Learning with Differential Privacy". In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. Vienna, Austria: Association for Computing Machinery, 2016, pp. 308–318. ISBN: 9781450341394.

[2] M. S. M. S. Annamalai, B. Balle, J. Hayes, G. Kaissis, and E. D. Cristofaro. *The Hitchhiker's Guide to Efficient, End-to-End, and Tight DP Auditing*. 2025. arXiv: 2506.16666.

[3] H. H. Arcolezi. *LDP-Audit GitHub Repository*. 2024. URL: https://github.com/hharcolezi/ldp-audit.

[4] H. H. Arcolezi and S. Gambs. "Revealing the True Cost of Locally Differentially Private Protocols: An Auditing Perspective". In: *Proceedings on Privacy Enhancing Technologies* 2024.4 (Oct. 2024), pp. 123–141.

[5] B. Balle, G. Cherubin, and J. Hayes. "Reconstructing Training Data with Informed Adversaries". In: *Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, 2022, pp. 1138–1156. DOI: 10.1109/SP46214.2022.9833677.

[6] B. Balle and Y.-X. Wang. "Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by J. Dy and A. Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 394–403. URL: https://proceedings.mlr.press/v80/balle18a.html.

[7] B. Bichsel, S. Steffen, I. Bogunovic, and M. Vechev. "DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers". In: *Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, 2021, pp. 391–409. DOI: 10.1109/SP40001.2021.00081.

[8] M. Bun, D. Desfontaines, C. Dwork, M. Naor, K. Nissim, A. Roth, A. Smith, T. Steinke, J. Ullman, and S. Vadhan. *Statistical inference is not a privacy violation*. https://differentialprivacy.org/inference-is-not-a-privacy-violation/. Accessed: 2025-06-10. June 2021.

[9] K. Chatzikokolakis, G. Cherubin, C. Palamidessi, and C. Troncoso. "Bayes security: A not so average metric". In: *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. IEEE. 2023, pp. 388–406.

[10] Z. Ding, Y. Wang, G. Wang, D. Zhang, and D. Kifer. "Detecting Violations of Differential Privacy". In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. CCS '18. Toronto, Canada: Association for Computing Machinery, 2018, pp. 475–489. ISBN: 9781450356930.

[11] J. Dong, A. Roth, and W. J. Su. *Gaussian Differential Privacy*. May 30, 2019. DOI: 10.48550/arXiv.1905.02383. arXiv: 1905.02383 [cs]. URL: http://arxiv.org/abs/1905.02383 (visited on 08/16/2025). Pre-published.

[12] C. Dwork. "Differential Privacy". In: *Automata, Languages and Programming*. Ed. by M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 1–12. ISBN: 9783540359081.

[13] C. Dwork, F. McSherry, K. Nissim, and A. Smith. "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. 2006, pp. 265–284. DOI: 10.1007/11681878_14.

[14] C. Dwork and A. Roth. "The Algorithmic Foundations of Differential Privacy". In: *Foundations and Trends in Theoretical Computer Science* 9.3–4 (2014), pp. 211–407.

[15] Ú. Erlingsson, I. Mironov, A. Raghunathan, and S. Song. "That which we call private". In: *ArXiv* abs/1908.03566 (2019).

[16] Ú. Erlingsson, V. Pihur, and A. Korolova. "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response". In: CCS '14. Scottsdale, Arizona, USA: Association for Computing Machinery, 2014, pp. 1054–1067. ISBN: 9781450329576.

[17] M. Fredrikson, S. Jha, and T. Ristenpart. "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1322–1333. ISBN: 9781450338325.

[18] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart. "Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing". In: *23rd USENIX Security Symposium (USENIX Security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 17–32. ISBN: 9781931971157.

[19] D. Gorla, L. Jalouzot, F. Granese, C. Palamidessi, and P. Piantanida. "On Estimating the Strength of Differentially Private Mechanisms in a Black-Box Setting". In: *IEEE Transactions on Dependable and Secure Computing* 22.5 (2025), pp. 5494–5507. DOI: 10.1109/TDSC.2025.3568160.

[20] P. Guerra-Balboa, A. Sauer, and T. Strufe. "Analysis and Measurement of Attack Resilience of Differential Privacy". In: *Proceedings of the 23rd Workshop on Privacy in the Electronic Society*. WPES '24. Salt Lake City, UT, USA: Association for Computing Machinery, 2024, pp. 155–171. ISBN: 9798400712395.

[21] M. E. Gursoy, L. Liu, K.-H. Chow, S. Truex, and W. Wei. "An Adversarial Approach to Protocol Analysis and Selection in Local Differential Privacy". In: *IEEE Transactions on Information Forensics and Security* 17 (2022), pp. 1785–1799.

[22] J. Hayes, B. Balle, and S. Mahloujifar. "Bounding training data reconstruction in DP-SGD". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 78696–78722.

[23] F. Houssiau, J. Jordon, S. N. Cohen, O. Daniel, A. Elliott, J. Geddes, C. Mole, C. Rangel-Smith, and L. Szpruch. *TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data*. 2022. arXiv: 2211.06550.

[24] T. Humphries, S. Oya, L. Tulloch, M. Rafuse, I. Goldberg, U. Hengartner, and F. Kerschbaum. "Investigating Membership Inference Attacks under Data Dependencies". In: *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. 2023, pp. 473–488.

[25] M. Jagielski, J. Ullman, and A. Oprea. "Auditing differentially private machine learning: How private is private sgd?" In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 22205–22216.

[26] B. Jayaraman. "Analyzing the Leaky Cauldron: Inference Attacks on Machine Learning". Ph.D. dissertation. University of Virginia, Dec. 2022. URL: https://libraetd.lib.virginia.edu/public_view/1r66j21378.

[27] B. Jayaraman. *EvaluatingDPML GitHub Repository*. 2022. URL: https://github.com/bargavj/EvaluatingDPML.

[28] B. Jayaraman and D. Evans. "Are Attribute Inference Attacks Just Imputation?" In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*. CCS '22. Los Angeles, CA, USA: Association for Computing Machinery, 2022, pp. 1569–1582. ISBN: 9781450394505.

[29] P. Kairouz, K. Bonawitz, and D. Ramage. "Discrete distribution estimation under local privacy". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 2436–2444.

[30] P. Kairouz, S. Oh, and P. Viswanath. "The Composition Theorem for Differential Privacy". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, July 2015, pp. 1376–1385. URL: https://proceedings.mlr.press/v37/kairouz15.html.

[31] D. Kifer, J. M. Abowd, R. Ashmead, R. Cumings-Menon, P. Leclerc, A. Machanavajjhala, W. Sexton, and P. Zhuravlev. "Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census". In: *arXiv preprint arXiv:2209.03310* (2022).

[32] B. Kulynych, J. F. Gomez, G. Kaissis, F. du Pin Calmon, and C. Troncoso. "Attack-aware noise calibration for differential privacy". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 134868–134901.

[33] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[34] J. Lee and C. Clifton. "How Much Is Enough? Choosing ϵ for Differential Privacy". In: *Information Security*. Ed. by X. Lai, J. Zhou, and H. Li. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 325–340. ISBN: 9783642248610.

[35] S. Lestyán, G. Ács, and G. Biczók. *In Search of Lost Utility: Private Location Data*. 2022. arXiv: 2008.01665.

[36] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc., 2017.

[37] Y. Lu, M. Magdon-Ismail, Y. Wei, and V. Zikas. "Eureka: A General Framework for Black-box Differential Privacy Estimators". In: *Symposium on Security and Privacy (SP)*. San Francisco, CA, USA: IEEE, 2024, pp. 913–931. DOI: 10.1109/SP54263.2024.00166.

[38] D. J. MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[39] S. Mahloujifar, L. Melis, and K. Chaudhuri. *Auditing f-Differential Privacy in One Run*. 2024. arXiv: 2410.22235.

[40] M. Malek Esmaeili, I. Mironov, K. Prasad, I. Shilov, and F. Tramer. "Antipodes of Label Differential Privacy: PATE and ALIBI". In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 6934–6945.

[41] S. Meiser. *Approximate and Probabilistic Differential Privacy Definitions*. Cryptology ePrint Archive, Paper 2018/277. 2018. URL: https://eprint.iacr.org/2018/277.

[42] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. "Unique in the Crowd: The privacy bounds of human mobility". In: *Scientific Reports* 3 (2013).

[43] M. S. Muthu Selva Annamalai and E. De Cristofaro. "Nearly Tight Black-Box Auditing of Differentially Private Machine Learning". In: *Advances in Neural Information Processing Systems*. Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang. Vol. 37. Curran Associates, Inc., 2024, pp. 131482–131502.

[44] P. Nanayakkara, M. A. Smart, R. Cummings, G. Kaptchuk, and E. M. Redmiles. "What are the chances? explaining the epsilon parameter in differential privacy". In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 1613–1630.

[45] A. Narayanan and V. Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (sp 2008)*. 2008, pp. 111–125. DOI: 10.1109/SP.2008.33.

[46] M. Nasr, S. Songi, A. Thakurta, N. Papernot, and N. Carlin. "Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning". In: *2021 IEEE Symposium on Security and Privacy (SP)*. 2021, pp. 866–882.

[47] M. O'Connell, M. Moreira, and W. Kan. *ECML/PKDD 15: Taxi Trajectory Prediction (I)*. Kaggle. 2015. URL: https://kaggle.com/competitions/pkdd-15-predict-taxi-service-trajectory-i.

[48] OpenStreetMap contributors. *Planet dump retrieved from https://planet.osm.org*. 2017. URL: https://www.openstreetmap.org.

[49] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro. "What Does The Crowd Say About You? Evaluating Aggregation-based Location Privacy". In: *Proceedings on Privacy Enhancing Technologies* 2017 (2017), pp. 156–176.

[50] T. Sauer. *Numerical analysis*. Pearson, 2018. ISBN: 978-0-321-78367-7.

[51] T. Steinke, M. Nasr, and M. Jagielski. "Privacy auditing with one (1) training run". In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS '23. New Orleans, LA, USA: Curran Associates Inc., 2023.

[52] L. Sweeney. *Simple Demographics Often Identify People Uniquely*. Data Privacy Working Paper 3. Carnegie Mellon University, Data Privacy Lab, 2000.

[53] TensorFlow contributors. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. 2025. URL: https://www.tensorflow.org.

[54] F. Tramèr, A. Terzis, T. Steinke, S. Song, M. Jagielski, and N. Carlini. "Debugging Differential Privacy: A Case Study for Privacy Auditing". In: *ArXiv* abs/2202.12219 (2022).

[55] T. Wang, J. Blocki, N. Li, and S. Jha. "Locally differentially private protocols for frequency estimation". In: *26th USENIX Security Symposium (USENIX Security 17)*. 2017, pp. 729–745.

[56] H. Xiao, K. Rasul, and R. Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. 2017. arXiv: 1708.07747.

[57] Y. Xiao and L. Xiong. "Protecting Locations with Differential Privacy under Temporal Correlations". In: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. CCS '15. Denver, Colorado, USA: Association for Computing Machinery, 2015, pp. 1298–1309. ISBN: 9781450338325.

[58] M. Ye and A. Barg. "Optimal Schemes for Discrete Distribution Estimation Under Locally Differential Privacy". In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5662–5676.

[59] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. "Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting". In: *2018 IEEE 31st Computer Security Foundations Symposium (CSF)* (2017), pp. 268–282.

[60] Y. Zheng, H. Fu, X. Xie, W.-Y. Ma, and Q. Li. *Geolife GPS trajectory dataset - User Guide*. Geolife GPS trajectories 1.1. July 2011. URL: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.