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) is widely adopted in data management systems to enable data sharing with formal disclosure guarantees. A central systems challenge is understanding how DP noise translates into effective protection against inference attacks, since this directly determines achievable utility. Most existing analyses focus only on membership inference–capturing only a threat–or rely on reconstruction robustness (ReRo). However, under realistic assumptions, we show that ReRo can yield misleading risk estimates and violate claimed bounds, limiting their usefulness for principled DP calibration and auditing.

This paper introduces reconstruction advantage, a unified risk metric that consistently captures risk across membership inference, attribute inference, and data reconstruction. We derive tight bounds that relate DP noise to adversarial advantage and characterize optimal adversarial strategies for arbitrary DP mechanisms and attacker knowledge. These results enable risk-driven noise calibration and provide a foundation for systematic DP auditing. We show that reconstruction advantage improves the accuracy and scope of DP auditing and enables more effective utility-privacy trade-offs in DP-enabled data management systems.

1. Introduction

Differential Privacy (DP) [24] and its distributed variant, local DP (LDP), have emerged as the de facto standard to mitigate privacy risk—that is, the extent to which a learning process allows sensitive information about participants to be inferred. DP aims to make participation as safe as not participating [22], and its privacy-utility trade-off is governed by the privacy budget ε (smaller values provide 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 [55]. Despite this solid theoretical foundation, a central practical question remains: How do these formal parameters, especially ε, translate into concrete protection against real-world attacks? [57] This question is critical for calibrating ε: if set too high, sensitive information may be exposed; if too low, utility is unnecessarily compromised. Furthermore, understanding this relationship is essential for DP auditing, which aims to empirically estimate privacy [38], test the tightness of DP mechanisms [59], and detect bugs [68].

Motivated by its applications in noise calibration and auditing, there is growing interest in the data management community in risk assessment for DP mechanisms [12, 16, 18, 34]. Significant progress has been made in connecting DP to the risk of *membership inference attacks* (MIAs) [12, 26, 37, 73], even enabling direct noise calibration for desired MIA risk levels [45] without explicitly choosing ε. However, MIAs capture only one aspect of privacy risk and may be less relevant in deployments such as census data releases. In particular, *attribute inference attacks* (AIAs) [73], which can expose sensitive information even when membership is public [9], remain less understood. Recently, *data reconstruction attacks* (DRAs) [9] were proposed as a unifying framework subsuming both MIAs and AIAs, while also accounting for partial or imperfect reconstruction, e.g., revealing a car's license plate may suffice to compromise 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 [9] introduced the first metric for DRAs, *reconstruction robustness* (ReRo), providing a pioneering unified view of DP attack resilience. ReRo was foundational, but has limitations as a comprehensive adversarial metric. First, ReRo and existing bounds [9, 33] assume attackers have no target-specific auxiliary knowledge, ignoring partial information such as demographic attributes or social media data—information that real-world attacks often exploit [56, 58, 66]. We empirically confirm this limitation: when target-specific auxiliary information is available, the empirical ReRo exceeds the existing ReRo bounds (see Figure 5). Second, ReRo is a success probability, which penalizes mechanisms for providing global statistical knowledge–the end goal of data release–and incorrectly accounts for success from background knowledge or statistical imputation as participation risk [15, 44], leading to unnecessary utility loss when used for noise calibration (Figure 2).

We address such limitations by introducing *reconstruction advantage* (RAD), which extends advantage metrics to the unifying DRA framework. RAD overcomes ReRo's limitations, naturally incorporating auxiliary knowledge and avoiding risk overestimation. We establish tight bounds linking DP noise to RAD, enabling noise injection calibrated to a participant's true risk of information disclosure. Specifically, we provide: (i) a worst-case bound independent of the attacker's auxiliary knowledge (Theorem 4.1), and (ii) an auxiliary-dependent, universally tight bound (Theorem 4.2). To assess tightness, we construct and prove the optimal attack strategy for any reconstruction goal, auxiliary knowledge, and mechanism–which also serves as a practical tool for DP auditing.

Theorem 4.2 is universally tight and cannot be further improved. However, it requires full knowledge of the mechanism $\mathcal{M}$, limiting its applicability in auditing external software. While Theorem 4.1 can serve as a fallback in such scenarios, it may strongly overestimate risk when no auxiliary information is available. To address this, we provide closed-form, black-box upper bounds for RAD without auxiliary knowledge (i.e., when the entire target record is considered secret, as in [6, 9, 33]) and for the case of perfect reconstruction, which is particularly relevant for categorical data where sensitive attributes (e.g., diseases, political opinions, or religious beliefs) cannot be partially reconstructed [27, 28]. All our bounds substantially reduce the required noise compared to existing ReRo bounds, and we validate these improvements experimentally.

These results provide the theoretical foundation for practical DP auditing. Modern DP systems deployed in industry [25], government [2], and data-processing pipelines [54] still lack general-purpose tools for quantifying real-world privacy leakage. Existing auditing tools either focus on a narrow attack class (often MIAs) [6, 38, 52, 59, 68] or rely on learning-based strategies requiring extensive tuning without mechanism-independent guarantees [50]. RAD fills this gap, offering a principled, mechanism-agnostic characterization of reconstruction risk. Building on our novel bounds, we introduce a RAD-based auditing framework that generalizes beyond prior tools [6, 20], capturing all reconstruction risks and providing more accurate, actionable privacy assessments. 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 [6]. 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 as demonstrated in our empirical study (see Figure 8).

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 applicability.
- We introduce *Reconstruction Advantage (RAD)* as a consistent, unifying risk metric that naturally incorporates auxiliary knowledge.
- We establish tight worst-case and auxiliary-dependent bounds for RAD, along with black-box bounds for attackers lacking auxiliary knowledge.
- We construct the optimal attack strategy for any reconstruction goal, mechanism, and prior distribution, proving its optimality and demonstrating empirical utility for auditing.

- We propose a RAD-based DP auditing framework that provides broader threat analyses and more accurate privacy-budget estimates than existing LDP auditing techniques.

This is the extended version of the paper under revision in the Proceedings of the VLDB Endowment (PVLDB), 2026. The code used for our experiments is 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.

2.1. 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 mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ which, given an input database $D \in \mathcal{Z}^n$, produces a global output (e.g., an aggregate statistic or a trained model) $\theta \in \Theta$ with probability/density function $p_{\mathcal{M}}(\theta \mid D)$. In this context, DP is formalized as follows:

Definition 2.1 ([24])**.** 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 [51].

If $\delta = 0$ we speak of *pure* DP (ε-DP). If $n = 1$, i.e., $\mathcal{M}$ takes as input a single data record $z \in \mathcal{Z}$, we obtain *Local Differential Privacy* (LDP). LDP is a rigorous and increasingly relevant privacy model in which data is randomized on the client side before being transmitted to a data collector [23]. Consequentially, it is especially suitable for privacy-sensitive applications such as telemetry and location-based services where no trusted data curator is considered [25].

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 [23]. 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.

Many real-world deployments apply multiple DP mechanisms sequentially [14, 19]. By DP's adaptative composition property, the total privacy loss is determined by the parameters of the individual mechanisms [43]. Formally, for each $i \in [T]$, let $\overline{\Theta}_{i-1} = \prod_{j=1}^{i-1} \Theta_j$ denote the space of previous outputs, and define $\mathcal{M}_i\colon \mathcal{Z}^n \times \overline{\Theta}_{i-1} \to \Theta_i$. The *$T$-fold composed* mechanism is $\mathcal{M}(D) = (\mathcal{M}_1(D), \mathcal{M}_2(D, Y_1), \ldots, \mathcal{M}_T(D, Y_{T-1}))$, where $Y_i = (\mathcal{M}_1(D), \ldots, \mathcal{M}_i(D, Y_{i-1}))$ denotes the first i outputs. Dwork et al. [24] established the first general bound on the privacy loss under T-fold adaptive composition: Composing (ε, δ)-DP mechanisms yields to $(T\varepsilon, T\delta)$-DP. Subsequent refinements led to the tighter composition bounds as presented in [43].

2.2. Differential Privacy and Attack Resilience.

Following previous work we consider for any target record z an *informed adversary* [9] with access to: the fixed dataset $D_- = D \setminus \{z\}$, 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 [9].

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 [9]: 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 the color of the car. Hence, DRAs cover the broad range of commonly discussed privacy risks, including MIAs and AIAs as a particular instance [9]. Formally, a DRA, denoted by $A\colon \Theta \times aux \to \mathcal{Z}$ uses the output of a DP mechanism $\theta \sim \mathcal{M}(D)$ and the target auxiliary information to produce a candidate $\tilde{z} = A(\theta, a(z))$. Note that, in case of composing several mechanisms, we consider the final output after the whole process.

The attack is considered successful if the output is similar enough (according to a success threshold η) to the real input z: $\ell(\tilde{z}, z) \leq \eta$. The error function ℓ depends on the context, for instance, in a classic AIA, given $z = (x, y)$ we define $\phi(z) = y$ and $\ell(\tilde{z}, z) = 0$ if $\phi(\tilde{z}) = \phi(z)$ and one otherwise. 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 sensitive enough to partially reconstruct the target, for instance, the image domain, even if not all pixels are correct. In this case, 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) [9]. Given the error function ℓ and the 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 [31, 73] agrees on the following metric:

Definition 2.2 (Adapted from [73])**.** 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 Definition 2.2 corrects for cases where the attribute could be inferred even without the record being in the database (e.g., through imputation [41]).

The current proposed performance metric for general DRAs [9] 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 [9])**.** Let π be a prior over $\mathcal{Z}$ and $\ell : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ a 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 $A : \Theta \to \mathcal{Z}$,

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

The first bound for ReRo under ε-DP was given by [9]:

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

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 [33] refined this bound using the f-DP [21], a characterization of DP that captures the exact statistical indistinguishability between neighbors through the functional f. Formally,

Definition 2.4 ([44])**.** 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 $D_0, D_1 \in \mathcal{Z}^n$ such that $d_H(D_0, D_1) \leq 1$ and all post-processing algorithms $A\colon \mathrm{Range}(\mathcal{M}) \to \mathcal{D}(\{0, 1\})$,

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

Here, f is known as a *trade-off function* [21], 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. When f is the trade-off function between two normal distributions with different means, namely $f(\alpha) = \Phi\big(\Phi^{-1}(1-\alpha) - \mu\big)$, where Φ denotes the standard normal CDF, the resulting notion is known as *Gaussian DP* (μ-GDP).

The f-DP framework facilitates the computation of quantities such as the *total variation* distance:

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 $\mathcal{M}$ satisfying (ε, δ)-DP, its TV is bounded [43] as

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

Both f-DP and TV are preserved under composition. Specifically, the T-fold composition of an f-DP mechanism satisfies $f^{\otimes T}$-DP, where $f \otimes f$ denotes the trade-off function $T(P \times P, Q \times Q)$ for $f = T(P, Q)$. For instance, if a mechanism is μ-GDP, then its T-fold composition is $(\mu\sqrt{T})$-GDP [21]. Moreover, if $\mathrm{TV}(\mathcal{M}_i) = \Delta$, then the T-fold composition satisfies $\mathrm{TV}(\mathcal{M}) \leq 1 - (1-\Delta)^T$ [29]. This bound can be sharpened to $\max_\alpha(1 - f^{\otimes T}(\alpha) - \alpha)$ when f is known.

Hayes, Balle, and Mahloujifar [33] present the first 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, the most known DP algorithm for private learning [1].

2.3. **Measure Theory Results.** In continuous probability spaces, events of the form $X = x$ have probability zero, so conditional probabilities defined via ratios are not well-defined. The disintegration theorem provides a rigorous substitute: any joint probability measure μ on $X \times Y$ can be decomposed as

$$\mu(dy\, dx) = \mu_x(dy)\, \mu_X(dx),$$

where μ_X is the marginal of X and μ_x is a probability measure on Y representing the conditional law of Y given $X = x$. This decomposition allows conditional distributions to be defined pointwise (almost everywhere), despite conditioning on null events.

This intuition extends from Cartesian products to general measurable maps $a\colon Z \to X$, where disintegration allows one to define conditional measures μ_x supported on the fibers $a^{-1}(x)$, providing a rigorous notion of conditioning on $a(Z) = x$ even when $\mu(a^{-1}(x)) = 0$. Formally:

Theorem 2.1 (Disintegration Theorem [8])**.** *Let $(Z, \mathcal{Z}, \mu)$ be a probability space, let $(X, \mathcal{X})$ be a standard Borel space, and let $a : Z \to X$ be a measurable map. Denote by $\nu = \mu \circ a^{-1}$ the push-forward measure of μ through a. Then there exists a ν-almost everywhere uniquely determined family of probability measures $\{\mu_x\}_{x \in X}$ on $(Z, \mathcal{Z})$ such that:*

(1) For ν-a.e. $x \in X$, μ_x is supported on the fiber $a^{-1}(x)$, i.e.

$$\mu_x\Big(Z \setminus a^{-1}(x)\Big) = 0.$$

(2) For every measurable set $B \in \mathcal{Z}$,

$$\mu(B) = \int_X \mu_x(B)\, d\nu(x).$$

(3) For every integrable function $f \in L^1(Z, \mu)$,

$$\int_Z f(z)\, d\mu(z) = \int_X \left(\int_{a^{-1}(x)} f(z)\, d\mu_x(z) \right) d\nu(x).$$

Note that when μ is a product measure $\mu_X \otimes \mu_Y$ and a is the projection onto X, the conditional measures μ_x can be taken equal to μ_Y for ν-almost every x, and the above reduces to the classical Fubini–Tonelli theorem.

Remark 1. The disintegration theorem applies straightforwardly when μ is the counting measure on a discrete space. Let Z and X be discrete sets and let $a : Z \to X$ be any function.

For each $x \in X$, define the fiber

$$Z_x := a^{-1}(x) = \{z \in Z : a(z) = x\}.$$

Let μ be the counting measure on Z, i.e., $\mu(B) = \#B$ for $B \subseteq Z$. The pushforward measure on X is then

$$\nu(\{x\}) = \mu(a^{-1}(x)) = \#Z_x.$$

For each x with $\nu(\{x\}) > 0$, define the conditional measure μ_x on Z by

$$\mu_x(B) = \frac{\#(B \cap Z_x)}{\#Z_x},$$

i.e., the uniform distribution on the fiber Z_x. Then for any $B \subseteq Z$,

$$\mu(B) = \sum_{x \in X} \mu_x(B)\, \nu(\{x\}).$$

Consequently, for any function $f : Z \to [0, \infty]$,

$$\sum_{z \in Z} f(z) = \sum_{x \in X} \left(\int f(z)\, \mu_x(dz) \right) \nu(\{x\}) = \sum_{x \in X} \left(\sum_{z \in Z_x} f(z) \right).$$

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, discussing novel insights on gaps that motivate our work.

3.1. Attack-Based DP Noise Calibration.

Several recent studies [12, 17, 45] demonstrate that calibrating DP noise based on resilience to specific attacks can significantly help improve utility. Such approaches, however, primarily target MIAs, which leads to unnecessary utility degradation without offering meaningful privacy benefits when membership is public or considered non-sensitive [9].

Beyond MIAs, privacy concerns often involve AIA, where the adversary aims to infer sensitive attributes of individuals from released data [39, 62]. A common metric for evaluating such attacks is the attribute advantage [73]. Existing works that provide theoretical bounds for AIAs either analyze specific attack strategies [73] or adopt more general DRA frameworks [9, 31]. 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 [9]. Moreover, Equation (1) [9] and Equation (3) [33] provide ReRo-based DP noise calibration methods.

3.2. A note on Limitations of ReRo.

A general-purpose risk metric would be expected to cover all relevant attack scenarios. 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 2.2.

Formally, the attack A in [9], corresponding to our Definition 2.3, 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 attacker A cannot adapt its strategy to a specific target z. This choice fundamentally prevents assessing the risk of MIA and AIA, as they use full or partial knowledge of some target records. This is a relevant limitation since most real-world privacy attacks historically exploit publicly available information about the target [56, 58, 66]. Moreover, we 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 exclude such knowledge, but all succeeding formal bounds connecting ReRo and DP were also proven under this restrictive exclusion. The requirement that the attack depends only on $\mathcal{M}(D)$—ignoring target-specific information—is critical to establishing both Equations (1) and (3) above. This is not merely a theoretical limitation: we show in Section 7 that these bounds do not hold for attacks that exploit target-specific knowledge against well-known mechanisms such as 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 [41], 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, ReRo still overestimates risk, as we discussed in our preliminary work [31]. 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* [22, 44]. Indeed, several works establish that it is impossible to simultaneously provide utility and eliminate absolute information gain [22, 44].

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 for a novel framework to more accurately assess the risk of DP mechanisms with respect to privacy attacks.

3.3. **DP Auditing.** DP auditing [4] seeks to demonstrate tight estimates of the privacy budget, discover implementation flaws, and estimate empirical privacy. However, auditing in practice remains a significant challenge. For instance, implementation bugs or design flaws can severely degrade privacy guarantees in ways that are not immediately obvious. To address this, black-box discovery methods such as DP-Sniper [13] and Eureka [50] have been developed to detect DP violations by training classifiers to distinguish between mechanism outputs from "worst-case" adjacent inputs. While effective at uncovering certain classes of violations, this assumption breaks down for frequency-oracle mechanisms over high-dimensional categorical domains, where outputs are discrete randomized encoding [7] with inherently combinatorial structure. Consequently, the learned classifiers fail to scale, becoming prohibitively slow or ineffective as the domain dimension grows.

Beyond identifying bugs, existing empirical privacy auditing approaches primarily focus on MIAs [3, 12, 38, 65], 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 [53] or synthetic data generation [36]. In the LDP setting, the state-of-the-art framework LDP Auditor [6] relies specifically on perfect reconstruction without target-specific auxiliary knowledge for auditing.

Summarizing, despite its practical importance, no existing auditing framework incorporates auxiliary information or supports a DRA-based analysis that goes beyond MIAs and enables systematic evaluation across diverse DP mechanisms. Our preliminary work [31] made partial progress by analyzing adversaries that rely solely on the mechanism output; however, it did not account for the impact of target-specific auxiliary information, which is often decisive in real-world privacy breaches, such as the classical census re-identification [66]. This gap motivates

the development of RAD, a general auditing methodology designed to capture realistic adversaries and to quantify broader classes of privacy risks.

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 prove its tightness by constructing the corresponding optimal attack that achieves it. Together, these results provide a noise calibration method to optimize utility for a given risk. We empirically validate the practical tightness of our bounds in 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 RAD. 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 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

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

RAD explicitly accounts for target-specific auxiliary knowledge, providing a generalization of the membership and attribute advantages to arbitrary reconstruction attacks. Importantly, 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, analogously to membership and attribute advantage [73]. 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 DRAs 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 [9, 31], which consequently fail for attacks utilizing target-specific auxiliary knowledge as demonstrated 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 [31], initially calling it U-ReRo, which, however, similar to ReRo, fails to take *aux* into account.

Theorem 4.1 ((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 [49, 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 Definition 2.5 to RAD Definition 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, **??** completes the result. □

Note that in the discrete case, $\kappa_\pi = \sum_z \pi_z^2$, hence, the worst-case prior corresponds $\pi = U\{z_0, z_1\}$. In the continuous case, this result simplifies to η-RAD$\leq \mathrm{TV}(\mathcal{M})$, unaffected by the prior distribution.

Theorem 4.1 is the first bound for RAD under the strongest threat model, where the attacker may use auxiliary knowledge. Experiments on real datasets (see Section 7) show that this bound is tight: attacks can achieve the predicted advantage, confirming that it accurately captures the worst-case scenario. Hence, it is a crucial tool for DP noise calibration, improving over ReRo.

Moreover, Theorem 4.1 allows upper bounding RAD under composition. Given $\mathrm{TV}(\mathcal{M}_i) = \Delta$, the T-fold adaptative composition satisfies $\mathrm{TV}(M) \leq (1-(1-\Delta)^T)$. Hence, η-RAD$\leq (1-(1-\Delta)^T)(1-\kappa_\pi)$.

Since Theorem 4.1 does not depend on the attacker's auxiliary knowledge, the same bound 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 ($a(z) = z$)—as assumed for MIA. Therefore, we next provide a tighter bound that explicitly incorporates the target-specific auxiliary knowledge.

Theorem 4.2. *Given $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ and $a\colon \mathcal{Z} \to aux$ measurable, then for any attack $A\colon \Theta \times aux \to \mathcal{D}(\mathcal{Z})$, we have*

$$\eta\text{-RAD} \leq \int_\Theta \int_{aux} \max_{z_\theta \in \mathcal{Z}} \left(\int_{S_\eta^x(z_\theta)} w(\theta, z)\pi_z \, \mathrm{d}\mu_x(z) \right) \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta)$$

where μ the counting/Lebesgue measure and π_z mass/density function in the discrete/continuous case, $w(z,\theta) = p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)$, $S_\eta^x(z_\theta) = \{z\colon a(z) = x \wedge \ell(z_\theta, z) \leq \eta\}$, $\nu(x) = \mu \circ a^{-1}(x)$ and μ_x the disintegration theorem measure. The discrete case simplifies to

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

by direct application of Remark 1.

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

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

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

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

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

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

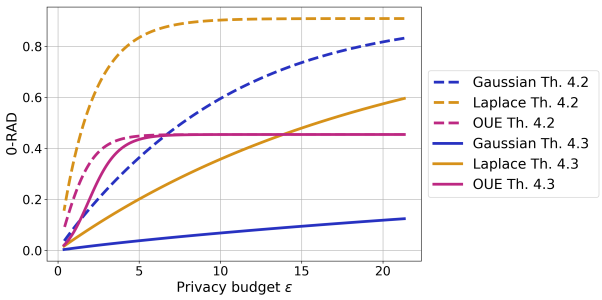


FIGURE 1. Improvement of Theorem 4.2 over Theorem 4.1 for different DP mechanisms, $|\mathcal{Z}| = 11$ and uniform prior.

Hence, given $a^{-1}(x) = \{z\colon a(z) = x\}$ for all $x \in aux$, given $\nu(x) = \mu \circ a^{-1}(x)$, applying disintegration theorem [8] it exists a unique measure μ_x such that

$$\begin{aligned}
\eta\text{-RAD} &= \int_{\mathcal{Z}} \int_{\Theta} \Pr_{A}(S_\eta(z) \mid a(z), \theta) w(z,\theta) \pi_z \,\mathrm{d}\mu(z)\,\mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(x)} \Pr_{A}(S_\eta(z) \mid x, \theta) w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(x)} \int_{\mathcal{Z}} \mathbf{1}_{\{\ell(z,\tilde{z}) \le \eta\}} p_A(\tilde{z} \mid x, \theta) w(z,\theta) \pi_z \,\mathrm{d}\mu(\tilde{z})\,\mathrm{d}\mu_x(z)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{\mathcal{Z}} p_A(\tilde{z} \mid x, \theta) \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(z,z') \le \eta\}} w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z)\,\mathrm{d}\mu(z\tilde{z})\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&\le \int_{\Theta} \int_{aux} \int_{\mathcal{Z}} p_A(\tilde{z} \mid x, \theta) \max_{z_\theta \in Z} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(z,z_\theta) \le \eta\}} w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z)\,\mathrm{d}\mu(\tilde{z})\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \max_{z_\theta \in Z} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(z,z_\theta) \le \eta\}} w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z) \left(\int_{\mathcal{Z}} p_A(\tilde{z} \mid x, \theta)\,\mathrm{d}\mu(\tilde{z}) \right) \mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \max_{z_\theta \in Z} \int_{a^{-1}(x) \cap \{z\colon \ell(z,z_\theta) \le \eta\}} w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta) \\
&\qquad \int_{\Theta} \int_{aux} \max_{z_\theta \in Z} \int_{S^x_\eta(z_\theta)} w(z,\theta) \pi_z \,\mathrm{d}\mu_x(z)\,\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta)
\end{aligned} \tag{15}$$

where $S^x_\eta(z_\theta) = \{z\colon a(z) = x \wedge \ell(z_\theta, z) \le \eta\}$. □

Theorem 4.2 bounds RAD when the specific $\mathcal{M}$ and auxiliary knowledge, aux, are known. At the same time, it becomes more precise than our worst-case bound Theorem 4.1. Moreover, it admits simple characterizations for commonly studied threat models. Particularly, if in a MIA, i.e., $a(z) = z$ for all records, $\nu(x) = \mu \circ a^{-1}(x) = \mu(x)$ and

$$\mu_x(\mathcal{Z} \backslash a^{-1}(x)) = \mu_x(\mathcal{Z} \backslash \{x\}) = 0 \Rightarrow \mu_x(z) = \delta_x(z),$$

where $\delta_x(z) = 1$ if $x = z$ and zero otherwise, therefore satisfying

$$\int_{aux} \delta_x(B)\,\mathrm{d}\nu(x) = \int_{aux} \mathbf{1}_{\{x \in B\}}\,\mathrm{d}\mu(x) = \mu(B).$$

Then, $\nu(z) = \mu(z)$, and $\mu_z(z) = \delta_z$. Hence,

$$\eta\text{-RAD} \le \int_{\Theta} \int_{aux} \max_{z_\theta \in \mathcal{Z}} \int_{\{x\}} \mathbf{1}_{\{\ell(z_\theta,z) \le \eta\}} w(\theta,z) \pi_z \,\mathrm{d}\delta_x(z)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(\theta) \tag{16}$$

$$\int_{\Theta} \int_{aux} \max_{z_\theta} \mathbf{1}_{\{\ell(z_\theta,x) \le \eta\}} w(\theta,x) \pi_x \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(\theta), \tag{17}$$

$$= \int_{\Theta} \int_{\{z\colon w(\theta,z) > 0\}} w(\theta,z) \pi_z \,\mathrm{d}\mu(\theta)\,\mathrm{d}\mu(z), \tag{18}$$

since $\arg\max_{z_\theta} = z$ if $w(\theta, z) > 0$ and $\arg\max_{z_\theta} = \mathcal{Z}\backslash S_\eta(z)$ otherwise. For discrete variable previous formula simplifies to

$$\eta\text{-RAD} \leq \sum_{\Theta\in\mathcal{Z}} \sum_{z\colon w(\theta,z)>0} w(\theta, z)\pi_z. \tag{19}$$

On the other extreme, when $aux = \{\varnothing\}$, $\nu(\varnothing) = \mu(\mathcal{Z}) = 1$, hence $\nu(\varnothing)$ is the Dirac measure $\mu = \delta_\varnothing(x)$. The first condition defining $\mu_\varnothing$ is then

$$\mu_\varnothing(\mathcal{Z}\backslash a^{-1}(\varnothing)) = \mu_\varnothing(\mathcal{Z}\backslash\mathcal{Z}) = \mu_\varnothing(\varnothing) = 0$$

which is satisfied by any measure by definition. Hence, we look to the second defining condition,

$$\mu(B) = \int_\varnothing \mu_\varnothing(B)\,\mathrm{d}\delta_\varnothing(\varnothing) = \mu_\varnothing(B), \tag{20}$$

hence $\mu_\varnothing = \mu$ obtaining:

$$\begin{aligned}\eta\text{-RAD} &\leq \int_\Theta \int_{aux} \max_{z_\theta\in\mathcal{Z}} \left(\int_{S_\eta^x(z_\theta)} w(\theta, z)\pi_z\,\mathrm{d}\mu_x(z)\right)\mathrm{d}\nu(x)\,\mathrm{d}\mu(\theta)\\ &= \int_\Theta \int_{\{\varnothing\}} \max_{z_\theta\in\mathcal{Z}} \left(\int_{S_\eta^x(z_\theta)} w(\theta, z)\pi_z\,\mathrm{d}\mu_\varnothing(z)\right)\mathrm{d}\delta_\varnothing(x)\,\mathrm{d}\mu(\theta)\\ &= \int_\Theta \max_{z_\theta\in\mathcal{Z}} \left(\int_{S_\eta^\varnothing(z_\theta)} w(\theta, z)\pi_z\,\mathrm{d}\mu(z)\right)\mathrm{d}\mu(\theta).\end{aligned}$$

Note that, $S_\eta^\varnothing(z_\theta) = \{z \in a^{-1}(\varnothing)\colon \ell(z_\theta, z) \leq \eta\} = \{z \in \mathcal{Z}\colon \ell(z_\theta, z) \leq \eta\}$, which simplifies for the discrete case to:

$$\eta\text{-RAD} \leq \sum_{\theta\in\Theta} \max_{z'\in\mathcal{Z}} \sum_{\ell(z',z)\leq\eta} w(\theta, z)\pi_z. \tag{21}$$

Moreover, if $\eta = 0$ (perfect reconstruction), such us any AIA setting and the original ReRo setting [33]), Theorem 4.2 formula simplifies to:

$$0\text{-RAD} \leq \sum_{\theta\in\Theta} \sum_{x\in aux} \max_{\substack{a(z)=x\\ w(z,\theta)>0}} w(z, \theta)\pi_z. \tag{22}$$

Importantly, 0-RAD is consistently zero for continuous random variables by definition.

Finally, given $|\mathcal{Z}| = m$ and $aux = \{\varnothing\}$, previous equation admits the simplification

$$0\text{-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i)\right)\pi_z, \tag{23}$$

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

$$\Theta_{i+1} = \{\theta \in \Theta\colon z_{i+1} \in \arg\max_j w(\theta, z_j)\pi_j\}\backslash \cup_{k=1}^{i} \Theta_k.$$

We illustrate the benefits of Theorem 4.2 on relevant DP mechanisms through next examples and visualizations in Figures 1 and 7.

Example 1. The generalized randomized response mechanism (GRR) [42] is an 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 $q = (\mathrm{e}^\varepsilon + m - 1)^{-1}$. Since, $p \geq q$ for all $\varepsilon \geq 0$,

$$w(\theta, z) = \begin{cases}(p - q)(1 - \pi_\theta) & \text{if } z = \theta\\ (q - p)\pi_\theta & \text{otherwise,}\end{cases} \tag{24}$$

and $w(z, \theta) > 0$ iff $z = \theta$. Hence, applying Theorem 4.2 for $a(z) = z$:

$$\eta\text{-RAD} = \sum_\theta (p - q)(1 - \pi_\theta)\pi_\theta = \frac{\mathrm{e}^\varepsilon - 1}{\mathrm{e}^\varepsilon + m - 1}(1 - \kappa_\pi) = \mathrm{TV}(1 - \kappa_\pi).$$

While, if we consider $aux = \{\varnothing\}$,

$$\eta\text{-RAD} = (p-q)(1 - \sum_{\theta} \pi_\theta \inf_{\ell(z_\theta,\theta)\le\eta} \Pr_{Z\sim\pi}[\ell(Z, z_\theta) \le \eta]).$$

Example 2 (OUE)**.** In the optimal unary encoding (OUE) mechanism [69] 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 $k_\theta \ge 1$

$$\Pr(\theta, z) = \begin{cases} P \equiv \frac{1}{2} q^{k_\theta - 1} p^{m-k_\theta} & \text{if } \theta_z = 1 \\ Q \equiv \frac{1}{2} q^{k} p^{m-k_\theta - 1} & \text{if } \theta_z \ne 1 \end{cases} \tag{25}$$

and $\Pr(\vec{0}, z) = \frac{1}{2}p^{m-1}$.

Hence, $\Pr(\vec{0}) = \frac{1}{2}p^{m-1}$ and $w(\vec{0}, z) = 0$ for all z. For all $\theta \ne \vec{0}$ we obtain,

$$p(\theta) = \frac{1}{2} q^{k_\theta - 1} p^{m-k_\theta} \underbrace{\left(\sum_{z:\,\theta_z=1} \pi_z\right)}_{S_\theta} + \frac{1}{2} q^{k} p^{m-k_\theta-1} \left(1 - \sum_{z:\,\theta_z=1} \pi_z\right)$$

Note that $P - Q = \frac{p-q}{2}(q^{k_\theta-1}p^{m-k_\theta-1}) \ge 0$. Consequently,

$$w(\theta, z) = \begin{cases} (P-Q)\,(1 - S_\theta) \ge 0 & \text{if } \theta_z = 1 \\ (Q-P) S_\theta \le 0 & \text{otherwise.} \end{cases}$$

Applying Theorem 4.2 for $a(z) = z$ we obtain,

$$\begin{aligned}
\eta\text{-RAD} &\le \sum_{\theta} \sum_{z:\theta_z=1} (P-Q)\,(1 - \sum_{\theta_z=1} \pi_z)\pi_z \\
&= \frac{p-q}{2} \sum_{\theta} \sum_{z:\theta_z=1} (q^{k_\theta-1}p^{m-k_\theta-1})\,(1 - \sum_{\theta_z=1} \pi_z)\pi_z \\
&= \frac{p-q}{2p} \sum_{k=1}^{m} q^{k-1}p^{m-k} \sum_{\theta:k_\theta=k} \sum_{\theta_z=1} \pi_z \Big(1 - \sum_{\theta_z=1} \pi_z\Big) \\
&= \frac{p-q}{2p} \sum_{k=1}^{m} q^{k-1}p^{m-k} \binom{m-2}{k-1} (1 - \kappa_\pi) \\
&= (1-\kappa_\pi) \frac{p-q}{2p} \sum_{k=1}^{m} \binom{m-2}{k-1} q^{k-1}p^{m-k} \\
&= (1-\kappa_\pi) \frac{p-q}{2p} \sum_{r=0}^{m-2} \binom{m-2}{r} q^{r}p^{m-1-r} \quad \text{(index change } r = k-1\text{)} \\
&= (1-\kappa_\pi) \frac{p-q}{2p}\, p^{m-1} \sum_{r=0}^{m-2} \binom{m-2}{r} (q/p)^{r} \\
&= (1-\kappa_\pi) \frac{p-q}{2p}\, p^{m-1} (1+q/p)^{m-2} \quad \text{(binomial identity)} \\
&= (1-\kappa_\pi) \frac{p-q}{2p}\, p^{m-1} (1/p)^{m-2} \quad \text{(since } p+q=1\text{)} \\
&= (1-\kappa_\pi) \frac{p-q}{2} \\
&= \frac{1}{2} \frac{e^\varepsilon - 1}{e^\varepsilon + 1} (1-\kappa_\pi) = \mathrm{TV}(\mathcal{M})\,(1-\kappa_\pi),
\end{aligned}$$

When $aux = \varnothing$, we have

$$\max_{z\in\mathcal{Z}} w(\theta,z)\pi_z = (P-Q)(1-\sum_{\theta_z=1}\pi_z)\max_{\theta_z=1}\pi_z.$$

Hence,

$$\begin{aligned}\eta\text{-RAD} &\le \sum_\theta (P-Q)\,(1-\sum_{\theta_z=1}\pi_z)\max_{\theta_z=1}\pi_z \\ &= \frac{p-q}{2}\sum_\theta (q^{k_\theta-1}p^{m-k_\theta-1})\,(1-\sum_{\theta_z=1}\pi_z)\max_{\theta_z=1}\pi_z\end{aligned}$$

We order $\pi_1 \le \cdots \le \pi_m$. Then,

$$\Theta_i = \{\theta\colon \max_z w(\theta,z)\pi_z = w(\theta,z_i)\} = \{\theta\colon \theta_i = 1 \wedge \theta_j = 0 \text{ for all } j > i\},$$

and we can rewrite

$$\eta\text{-RAD} \le \frac{p-q}{2p}\sum_{i=1}^m \pi_i \underbrace{\sum_{\theta\in\Theta_i} q^{k_\theta-1}p^{m-k_\theta}(1-\sum_{\theta_z=1}\pi_z)}_{A_i}$$

For every $k = 1,\dots,i$, there are $\binom{i-1}{k-1}$ vectors $\theta\in\Theta_i$ such that $k_\theta = k$. Moreover, the sum $\sum_{z\in S}\pi_z$ over all sets S of size k containing i:

$$\sum_{\substack{S\subseteq\{1,\dots,i\}\\ i\in S,|S|=k}}\sum_{z\in S}\pi_z = \pi_i\binom{i-1}{k-1} + \sum_{z=1}^{i-1}\pi_z\binom{i-2}{k-2}.$$

Hence,

$$\begin{aligned}A_i =& \sum_{k=1}^i q^{k-1}p^{m-k}\Big[\binom{i-1}{k-1}(1-\pi_i) - \binom{i-2}{k-2}\sum_{z=1}^{i-1}\pi_z\Big] \\ =&(1-\pi_i)\sum_{k=1}^i\binom{i-1}{k-1}q^{k-1}p^{m-k} - \Big(\sum_{z=1}^{i-1}\pi_z\Big)\sum_{k=1}^i\binom{i-2}{k-2}q^{k-1}p^{m-k} \quad \text{(index change } r=k-1,\ j=k-2\text{)} \\ =&(1-\pi_i)p^{m-i}\sum_{r=0}^{i-1}\binom{i-1}{r}q^r p^{i-1-r} - \Big(\sum_{z=1}^{i-1}\pi_z\Big)qp^{m-i}\sum_{j=0}^{i-2}\binom{i-2}{j}q^j p^{(i-2-j)} \\ =&(1-\pi_i)p^{m-i} + qp^{m-i}\left(\sum_{z=1}^{i-1}\pi_z\right).\end{aligned}$$

since for $k=1$, $\binom{i-2}{k-1} = 0$, so we can start in 2, i.e., $k = j+2$. Hence,

$$A_i = p^{m-i}\Big[(1-\pi_i) - q\sum_{z=1}^{i-1}\pi_z\Big], \qquad i = 1,\dots,m.$$

so

$$\eta\text{-RAD} \le \frac{p-q}{2p}\left(\sum_{i=1}^m p^{m-i}\pi_i(1-\pi_i) - q\sum_{i=1}^m p^{m-i}\pi_i\sum_{z=1}^{i-1}\pi_z\right)$$

For instance, $\pi_i = \frac{1}{m}$

$$\begin{aligned}\eta\text{-RAD} &\le \frac{p-q}{2p}\sum_{i=1}^m p^{m-i}\pi_i(1-\pi_i) - q\sum_{i=1}^m p^{m-i}\pi_i\sum_{z=1}^{i-1}\pi_z \\ &= \frac{p-q}{2mp}\left(\frac{m-1}{m}\sum_{i=1}^m p^{m-i} - q\frac{1}{m}\sum_{i=1}^m p^{m-i}(1-i)\right)\end{aligned}$$

$$= \frac{p-q}{2p}\left[\frac{1}{m}\Big(1-\frac{1}{m}\Big)\frac{p^m-1}{p-1} - q\cdot\frac{1}{m^2}\cdot\frac{p^m-1-m(p-1)}{(p-1)^2}\right]$$

$$\overset{p=1-q}{=} \frac{1-2q}{2(1-q)}\left[-\frac{m-1}{m^2}\cdot\frac{(1-q)^m-1}{q} - \frac{1}{m^2}\cdot\frac{(1-q)^m-1+mq}{q}\right]$$

$$= \frac{1-2q}{2(1-q)}\left(-\frac{1}{mq}((1-q)^m-1+q)\right)$$

$$= \frac{2q-1}{2(1-q)}\cdot\frac{(1-q)^m-1+q}{mq}$$

$$\overset{q=1-p}{=} \frac{1-2p}{2(1-p)p}\cdot\frac{p^m-p}{m} = \frac{1-2p}{2(1-p)}\cdot\frac{p^{m-1}-1}{m}$$

$$= \frac{(2p-1)(1-p^{m-1})}{2m(1-p)}.$$

Note that

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

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

Example 3 (SS, $aux = \varnothing$)**.** In the subset selection mechanism (SS) [72] users report a subset $\theta \subseteq \mathcal{Z} = \{z_1, \ldots, z_m\}$ containing their true value z with probability $p = \frac{\omega \mathrm{e}^\varepsilon}{\omega \mathrm{e}^\varepsilon + m - \omega}$, where $\omega = |\theta| = \max\left(1, \left\lfloor \frac{m}{\mathrm{e}^\varepsilon+1}\right\rfloor\right)$. The subset is completed by sampling uniformly from $\mathcal{Z}\setminus\{z\}$.

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

$$\Pr_{\mathcal{M}}(\theta \mid z) = \begin{cases} \frac{p}{A} & \text{if } z \in \theta \\ \frac{1-p}{B} & \text{if } z \notin \theta \end{cases} \tag{26}$$

Since $|\Theta| = \binom{m}{\omega}$ we have that, according to Equation (23), for $\pi = U[m]$,

$$\text{0-RAD} \le \frac{1}{m}\Big(\sum_{\theta\in\Theta}\max_z p_{\mathcal{M}}(\theta \mid z) - 1\Big) \tag{27}$$

$$= \frac{1}{m}\binom{m}{\omega}\frac{p}{A} = \frac{m}{m\omega}p - \frac{1}{m} = \frac{pm-\omega}{m\omega}. \tag{28}$$

Example 4 (Gaussian mechanism and $aux = \varnothing$)**.** The Gaussian mechanism adds Gaussian noise $\mathcal{N}(0,\sigma)$ the query value $q(D) \in \mathbb{R}$ [10]. If $\mathcal{Z} = \{z_1, \ldots, z_m\}$ is uniformly distributed and $\Delta q = 1$,

$$z \in \arg\max_j w(\theta, z_j)\pi_j \Leftrightarrow z \in \arg\max_j w(\theta, z_j).$$

Hence, applying Equation (23) we obtain

$$\text{0-RAD} \le \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - p_{\mathcal{M}}(\Theta_i)\right) = \frac{1}{m}\sum_{i=1}^{m}\left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - 1\right) \tag{29}$$

Note that for each z, $\Pr_{\mathcal{M}}(\theta \mid z) = \Pr_{\mathcal{M}}(\theta \mid q(D_z))$. 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:

$$\Theta_1 = (-\infty, \tfrac{z_1+z_2}{2}], \tag{30}$$

$$\Theta_i = [\tfrac{z_{i-1}+z_i}{2}, \tfrac{z_i+z_{i+1}}{2}], \quad 2 \le i \le n-1, \tag{31}$$

$$\Theta_n = [\tfrac{z_{n-1}+z_n}{2}, \infty). \tag{32}$$

On Θ_i, the maximizer is z_i. 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,

$$\text{0-RAD} \le \frac{2}{m}\sum_{j=1}^{m-1} \Phi\Big(\frac{\Delta_j}{2\sigma}\Big) \;-\; \frac{m-1}{m} \tag{33}$$

$$\le \frac{2(m-1)}{m}\Phi\Big(\frac{1}{(m-1)}\sum_{j=1}^{m-1}\frac{\Delta_j}{2\sigma}\Big) \;-\; \frac{m-1}{m} \tag{34}$$

$$\le \frac{m-1}{m}\left(2\Phi\Big(\frac{1}{2\sigma(m-1)}\Big) \;-1\right). \tag{35}$$

Where, Equation (34) follows since $\Delta_j \ge 0$, hence Φ concave, and we can apply Jensen's inequality, and Equation (35) since $\Delta q = 1$ therefore, $\sum_{j=1}^{m-1} \Delta_j = \Delta q = 1$.

Example 5 (Laplace Mechanism and $aux = \varnothing$)**.** The Laplace mechanism adds Laplace noise with scale $b = \Delta q/\varepsilon$ to the query value $q(D) \in \mathbb{R}$ [23]. If $\mathcal{Z} = \{z_1, \ldots, z_m\}$ if uniformly distributed and $\Delta q = 1$, analogously to Example 4,

$$z \in \arg\max_j w(\theta, z_j)\pi_j \Leftrightarrow z \in \arg\max_j w(\theta, z_j).$$

Hence, applying Equation (23) we obtain

$$\text{0-RAD} \le \frac{1}{m}\sum_{i=1}^{m}\Big(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - p_{\mathcal{M}}(\Theta_i)\Big) = \frac{1}{m}\sum_{i=1}^{m}\Big(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - 1\Big) \tag{36}$$

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 Example 4. Given the Laplace distribution CDF

$$F_i(x) = \begin{cases} \frac{1}{2}\exp\left(\frac{x-z_i}{b}\right) & \text{if } x < z_i \\ 1 - \frac{1}{2}\exp\left(-\frac{x-z_i}{b}\right) & \text{if } x \ge z_i \end{cases}, \tag{37}$$

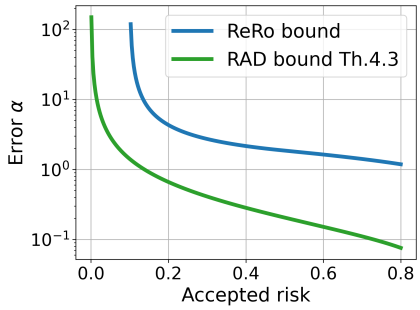


FIGURE 2. Upper bound on the Laplace mechanism query error (utility) at 95% confidence when the noise is calibrated using ReRo vs. RAD. We see that for the same risk estimation, calibrating with using RAD improves utility.

for $i = 1$,

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

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,

$$\sum_{i=1}^{m}\Pr_{\mathcal{M}}(\Theta_i \mid z_i) = m - \frac{1}{2}\left(\mathrm{e}^{-\frac{\varepsilon\Delta_1}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{m-1}}{2}} + \sum_{i=2}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_i}{2}} + \mathrm{e}^{-\frac{\varepsilon\Delta_{i-1}}{2}}\right) = m - \sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}} \tag{38}$$

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

$$\begin{aligned}
\text{0-RAD} \leq & \frac{1}{m}\left(m - 1 - \sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}}\right) && (39)\\
\leq & \frac{m-1}{m} - \frac{1}{m}\sum_{j=1}^{m-1}\mathrm{e}^{-\frac{\varepsilon\Delta_j}{2}} && (40)\\
\leq & \frac{m-1}{m} - \frac{m-1}{m}\mathrm{e}^{-\frac{1}{m-1}\sum_j\frac{\varepsilon\Delta_j}{2}} && (41)\\
\leq & \frac{m-1}{m}\left(1 - \mathrm{e}^{-\frac{\varepsilon}{2(m-1)}}\right). && (42)
\end{aligned}$$

Where, Equation (41) follows since $\Delta_j \geq 0$, hence Φ concave, and we can apply Jensen's inequality, and Equation (42) since $\Delta q = 1$ therefore, $\sum_{j=1}^{m-1}\Delta_j = \Delta q = 1$.

These examples demonstrate the applicability of Theorem 4.2 to estimate the risk in real-world scenarios. In Figure 1 we see the improvement when we target specific auxiliary knowledge instead of using our worst-case bound (Theorem 4.1). Hence, Theorem 4.2 offers an improved noise calibration method to ensure protection against real attacks, when the auxiliary knowledge is well defined. For instance, when the entire record is considered private ($aux = \{\varnothing\}$); alternatively, when a specific attribute y is deemed sensitive, and we consider all the remainder record public, i.e., $a(z) = z\backslash y$.

Importantly, we illustrate in Figure 2 the utility gain of noise calibration using our RAD bounds compared to using the best existing ReRo bound [33], showing the benefit of our bounds for system design. Specifically, we consider $aux = \{\varnothing\}$–allowing comparison with [33]. We plot

Algorithm 1: Optimal Attack

Input : θ and $a(z) = x$
Output : $\tilde{z}$
Compute $a^{-1}(x) = \{z \colon a(z) = x\}$
for $z' \in \mathcal{Z}$ **do**
  $\mathcal{W}_\eta^x(z') = \sum_{z \in a^{-1}(x) \colon \ell(z,z') \le \eta} w(\theta, z)\pi_z;$
Select $\tilde{z} \in \arg\max_{z'} \mathcal{W}_\eta^x(z')$ (at random)

the upper bound on the Laplace mechanism's query error that can be guaranteed with 95% confidence, for $|\mathcal{Z}| = 10$ and $\Delta = 1$, showing a substantial improvement in utility enabled by our RAD-based calibration.

Crucially, Theorem 4.2 is universally tight: for any mechanism and auxiliary knowledge, there exists an attack achieving the bound, so it cannot be further improved. We illustrate this by explicitly constructing such an attack in Algorithm 1, proving the existence of an optimal adversary for any auxiliary model. This result is particularly relevant to the database community, as it implies that, for a given risk tolerance, the utility of a mechanism cannot exceed what our method achieves; in other words, our approach yields optimal noise calibration.

Corollary 4.1 (Attack Optimality)**.** *Given the conditions as in Theorem 4.2, Algorithm 1 achieves the highest attainable η-RAD.*

Proof. Following Algorithm 1, given θ, x the attack outputs

$$\Pr_A(A(\theta, x) \in S(\theta, x)) = 1, \qquad p_A(a \mid \theta, x) = \frac{\mathbf{1}_{\{a \in S(\theta,x)\}}}{\mu(S(\theta, x))}.$$

with

$$S(\theta, x) = \arg\max_{a \in \mathcal{Z}} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(a,z_1) \le \eta\}}\, w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu_x(z_1). \tag{43}$$

Substituting in the definition of RAD and using the reformulation in Equation (11), we obtain

$$\begin{aligned}
\eta\text{-RAD} &= \int_{\mathcal{Z}} \int_{\Theta} \Pr_A\left(S_\eta(z_1) \mid \theta, a(z_1)\right) w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu(\theta)\, \mathrm{d}\mu(z_1) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(x)} \Pr_A\left(S_\eta(z_1) \mid \theta, x\right) w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu_x(z_1)\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{a^{-1}(x)} \int_{S(\theta,x)} \mathbf{1}_{\{\ell(a,z_1) \le \eta\}}\, \frac{1}{\mu(S(\theta, x))}\, w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu(a)\, \mathrm{d}\mu_x(z_1)\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \int_{S(\theta,x)} \frac{1}{\mu(S(\theta, x))} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(a,z_1) \le \eta\}}\, w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu_x(z_1)\, \mathrm{d}\mu(a)\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta).
\end{aligned}$$

Hence, applying Equation (43):

$$\begin{aligned}
\eta\text{-RAD} &= \int_{\Theta} \int_{aux} \int_{S(\theta,x)} \frac{1}{\mu(S(\theta, x))} \max_{a \in \mathcal{Z}} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(a,z_1) \le \eta\}}\, w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu_x(z_1)\, \mathrm{d}\mu(a)\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta) \\
&= \int_{\Theta} \int_{aux} \max_{a \in \mathcal{Z}} \int_{a^{-1}(x)} \mathbf{1}_{\{\ell(a,z_1) \le \eta\}}\, w(z_1, \theta)\, \pi_{z_1}\, \mathrm{d}\mu_x(z_1)\, \mathrm{d}\nu(x)\, \mathrm{d}\mu(\theta).
\end{aligned}$$

which according to Theorem 4.2, coincides with the maximun attainable bound. □

Corollary 4.1 directly establishes that Theorem 4.2 is universally tight and Theorem 4.1 is tight, since there exists at least one mechanism (GRR Example 1) for which Theorem 4.1 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 5c).

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 of our bounds in

(see Section 7). We also provide the application of our optimal attack in the specific case of DP-SGD:

Example 6 (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 [73], 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 [33], 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. Note that in each DP-SGD iteration $\bar{g}_t$ is obtained as $\sum_z \text{clip}_C(\nabla_\theta \ell(\theta_t, z)) + \mathcal{N}(0, c^2\sigma^2 I)$ where σ is the noise scale and $\text{clip}_C(\vec{v}) = \vec{v}\min(1, \frac{C}{\|\vec{v}\|_2})$ and θ_t released weights the previous iteration.

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:

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

and simplify w maximization to

$$\arg\max_{z:a(z)=x} w(\theta, z) = \arg\max_{z:a(z)=x} \sum_t W(g_t, \text{clip}_C(\nabla_\theta \ell(\theta_t, z))) \tag{44}$$

where $W(x, y) = \langle x, y\rangle - \frac{1}{m}\sum_z \langle x, \text{clip}_C(\nabla_\theta \ell(\theta_t, z))\rangle$, 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

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

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,

$$w(g, z) = \prod_t P_{\mathcal{M}}(g \mid z) - \prod_t P_{\mathcal{M}}(g_t) > 0 \Leftrightarrow \tag{45}$$

$$A^T\left(\prod_t \mathrm{e}^{B\langle g_t, \mu_z\rangle} - \prod_t \frac{1}{m}\sum_i \mathrm{e}^{B\langle g_t, \mu_{z_i}\rangle}\right) > 0 \Leftrightarrow \tag{46}$$

$$\mathrm{e}^{B\sum_t \langle g_t, \mu_z\rangle} > \prod_t \frac{1}{m}\sum_i \mathrm{e}^{B\langle g_t, \mu_{z_i}\rangle} \Leftrightarrow \tag{47}$$

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

$$B\sum_t \langle g_t, \mu_z\rangle > \sum_t \frac{1}{m}\sum_z \ln(\mathrm{e}^{B\langle g_t, \mu_{z_i}\rangle}) \Leftrightarrow \tag{49}$$

$$B\sum_t \langle g_t, \mu_z\rangle > \sum_t \frac{B}{m}\sum_i \langle g_t, \mu_{z_i}\rangle \Leftrightarrow \tag{50}$$

$$\sum_t \langle g_t, \mu_z\rangle - \sum_t \frac{1}{m}\sum_z \langle g_t, \mu_z\rangle > 0 \Leftrightarrow \tag{51}$$

$$\sum_t W(g_t, z) > 0. \tag{52}$$

Where Equation (49) 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 [33]. 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
  $\tilde{z} \longleftarrow U[\mathcal{Z} \backslash \{z\colon a(z) = x\}]$;

Our bounds offer concrete guidance for algorithm design, as they can be directly leveraged for noise calibration to achieve rigorous privacy guarantees while maximizing utility. In particular, they induce a simple protocol for practitioners. First, one must specify which information is deemed private (e.g., the full record, a subset of attributes, or membership), which determines the choice of the auxiliary information aux and $a\colon \mathcal{Z} \to aux$. Second, if prior knowledge about the distribution of $\mathcal{Z}$ is available, it should be encoded in a distribution π. If this is not the case, however, one must resort to the worst-case prior; otherwise, the attacker's risk may be underestimated. This worst-case prior typically corresponds to $\pi_x = \pi_y = 1/2$ for the two records that are easiest to distinguish (see Examples 1 and 2 and Figure 7). Nevertheless, even when the worst-case prior cannot be explicitly identified, the total variation bound given in Theorem 4.1 provides a safe upper bound for any choice of prior and auxiliary knowledge.

Third, the resulting RAD of the mechanism can be computed using Theorem 4.2–an auxiliary-dependent bound proven to be universally tight, or upper-bounded by a worst-case guarantee when the nature of aux is unknown (Theorem 4.1). Finally, by inverting the corresponding bound, one can directly derive the noise-injection parameters that meet a prescribed risk level. Since our bounds are tight, this procedure yields mechanisms that are utility-optimal for any given risk acceptance.

Note that while the closed form of Theorem 4.2 is easy to derive for discrete data, this may not hold for continuous data, where the bound involves Lebesgue integrals. In such case, the bound can be evaluated numerically using a nested Monte Carlo procedure. While numerical approximations introduce error, we show next how to obtain controlled confidence intervals in practice. As a safer alternative, one may always use our closed-form upper bound in Theorem 4.1.

However, this bound can be overly conservative when $aux = \{\varnothing\}$, motivating the tighter closed-form upper-bounds derived in the next section, which avoid numerical procedures even for continuous data.

4.1. **Numerical Approximation of Theorem 4.2.** We focus on approximating Theorem 4.2 in the continuous data domain $\mathcal{Z} = [a, b] \subseteq \mathbb{R}$. For simplicity, we consider the ℓ_1 metric, i.e., $\ell(z, z') = |z - z'|$, so that

$$S_\eta(z) = [\max(z - \eta, a), \min(z + \eta, b)] = [l(z), u(z)],$$

and we set $aux = \{\varnothing\}$ (otherwise one could directly use the closed-form in Theorem 4.1).

We assume $\mathcal{M}\colon \mathcal{Z} \to \mathcal{D}(\mathcal{Z})$ is a local differential privacy (LDP) mechanism. Given D_-, any global mechanism can be reduced to its local version $\mathcal{M}(z) = (\mathcal{M}(q(D_z)) - q(D_-)$, where q is the query. Hence, our goal is to approximate

$$\eta\text{-RAD} \leq \int_a^b \max_{x \in \mathcal{Z}} \int_{l(x)}^{u(x)} [p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)] \pi_z \, dz \, d\theta,$$

where π_z denotes the density function evaluated at z.

Outer integral: Monte Carlo approximation. The integral with respect to θ can be approximated using Monte Carlo integration, which is unbiased and has a controlled error via Hoeffding's inequality [35]. Hence, we can estimate RAD sampling $\theta_1, \ldots, \theta_{N_\theta} \sim U[a, b]$ and computing

$$\tilde{\gamma} = \frac{b - a}{N_\theta} \sum_{i=1}^{N_\theta} \underbrace{\max_{z_\theta \in \mathcal{Z}} \int_{l(x)}^{u(x)} [p_{\mathcal{M}}(\theta_i \mid z) - p_{\mathcal{M}}(\theta_i)] \pi_z \, dz}_{f(\theta_i)}.$$

Given $\gamma = \eta$-RAD and $f(\theta) \in [-M, M]$ over $[a, b]$, applying Hoeffding's inequality [35], we have

$$\Pr[|\tilde{\gamma} - \gamma| > t] \leq 2 \exp\Big(- \frac{N_\theta t^2}{M^2 (b - a)^2} \Big).$$

Note that $f(\theta) \in [-M, M]$ trivially with $M = \max_{z,\theta} p_{\mathcal{M}}(\theta \mid z)$. For instance, for the exponential mechanism with $u = -|z - \theta|$, we have

$$M = \max_{z,\theta} p_{\mathcal{M}}(\theta \mid z) = \frac{1}{s(1 - \exp(-1/s))}, \quad \text{where } s = \frac{2\Delta}{\varepsilon}.$$

Inner integral: Nested Monte Carlo approximation. If the integral with respect to z also requires a numerical approximation, we proceed with a nested Monte Carlo approach. To avoid bias, we first find z_θ analytically and then perform a nested Monte Carlo procedure.

Denote by z_θ any point achieving the maximum:

$$\max_{x \in \mathcal{Z}} \int_{l(x)}^{u(x)} [p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)] \pi_z \, dz \, d\theta.$$

To compute the z_θ for each θ, we use the Leibniz rule. For $\eta \leq (b - a)/2$:

$$S_\eta(x) = [l(x), u(x)] = \begin{cases} [a, x + \eta] & x \in [a, a + \eta], \\ [x - \eta, x + \eta] & x \in [a + \eta, b - \eta], \\ [x - \eta, b] & x \in [b - \eta, b]. \end{cases}$$

Let $g_\theta(x) = \int_{l(x)}^{u(x)} (p_{\mathcal{M}}(\theta \mid z) - p_{\mathcal{M}}(\theta)) \, \pi_z \, dz$. Then

- $x \in [a, a + \eta]$: $g'_\theta(x) = (p_{\mathcal{M}}(\theta \mid x + \eta) - p_{\mathcal{M}}(\theta)) \, \pi(x + \eta)$,
- $x \in [a + \eta, b - \eta]$: $g'_\theta(x) = (p_{\mathcal{M}}(\theta \mid x + \eta) - p_{\mathcal{M}}(\theta)) \pi(x + \eta) - (p_{\mathcal{M}}(\theta \mid x - \eta) - p_{\mathcal{M}}(\theta)) \pi(x - \eta)$,
- $x \in [b - \eta, b]$: $g'_\theta(x) = -(p_{\mathcal{M}}(\theta \mid x - \eta) - p_{\mathcal{M}}(\theta)) \pi(x - \eta)$.

The analogous follows for $(b-a)/2 \leq \eta \leq b$, but in such case

$$S_\eta(x) = [l(x), u(x)] = \begin{cases} [a, x+\eta] & x \in [a, b-\eta], \\ [a, b] & x \in [b-\eta, a+\eta], \\ [x-\eta, b] & x \in [a+\eta, b]. \end{cases}$$

Note that if $p_{\mathcal{M}}(\theta) = y$ is known, then $g'_\theta(x)$ can be evaluated explicitly. Consequently, z_θ can either be computed exactly or reduced to a finite set of candidate points. Since g is continuous and g'_θ is explicit, this set contains at most $k \leq 4$ candidates: the endpoints of the domain intervals and the critical points where $g'_\theta(x) = 0$, with $g'_\theta(x) \leq 0$ immediately before and $g'_\theta(x) \geq 0$ immediately after.

However, $p_{\mathcal{M}}(\theta)$ is itself an integral, $\mathbb{E}_{Z\sim\pi}[p_{\mathcal{M}}(\theta \mid Z)]$. In many cases, this integral may not have a close form, consequently, we also need to numerically approximate it. However, thanks to the Lipchitz properties of g respect to $p_{\mathcal{M}}(\theta)$ we can upper-bound the error as we will see in Equation (57).

First, we write

$$g(\theta, x, y) = \int_a^b \mathbf{1}_{\{|z-x|\leq\eta\}}(p_{\mathcal{M}}(\theta \mid z) - y)\pi_z \, \mathrm{d}z.$$

Note that g is Lipchitz with respect to y:

$$|g(\theta, x, \tilde{y}) - g(\theta, x, y)| = \int_a^b \mathbf{1}_{\{|z-x|\leq\eta\}}|\tilde{y} - y|\pi_z \, \mathrm{d}z \tag{53}$$

$$= |\tilde{y} - y| \int_a^b \mathbf{1}_{\{|z-x|\leq\eta\}}\pi_z \, \mathrm{d}z \leq \kappa^+_{\pi,\eta}|\tilde{y} - y|. \tag{54}$$

Importantly, it works uniformly for every $x \in [a, b]$. Hence, we can bound the error when estimating the maximum:

$$|\max_x g(\theta, x, \tilde{y}) - \max_x g(\theta, x, y)| \leq |\max_x(g(\theta, x, \tilde{y}) - g(\theta, x, y))| \leq \kappa^+|\tilde{y} - y|. \tag{55}$$

Given θ and y fixed, we denote $\mathcal{K}_{y,\theta}$ the candidates to maximum deduced using Leibniz. As mentioned $|\mathcal{K}| \leq 4$.

Now, we have three Monte Carlo procedures nested. From each of them we control de error, hence using the triangular inequality we can give a concentration upper-bound for the whole process.

We use the following notation (Table 1):

Target	**Estimator**
$y = p(\theta) = \int_a^b p_{\mathcal{M}}(\theta \mid z)\, \pi_z \, \mathrm{d}z$	$\hat{y} = \hat{p}(\theta) = \frac{1}{N_p}\sum_{i=1}^{N_p} p_{\mathcal{M}}(\theta \mid z_i), \quad z_1, \ldots, z_{N_p} \sim \pi$
$g(\theta, x, y) = \int_a^b \mathbf{1}_{\{\lvert z-x\rvert\leq\eta\}}(p_{\mathcal{M}}(\theta \mid z) - y)\, \pi_z \, \mathrm{d}z$	$\hat{g}(\theta, x, y) = \frac{1}{N_z}\sum_{i=1}^{N_z} \mathbf{1}_{\{\lvert z_i-x\rvert\leq\eta\}}(p_{\mathcal{M}}(\theta \mid z_i) - y),\ z_1, \ldots, z_{N_z} \sim \pi$
$f(\theta) = \max_{x\in\mathcal{K}_\theta} g(\theta, x, p(\theta))$	$\hat{f}(\theta) = \max_{x\in\mathcal{K}_\theta} \hat{g}(\theta, x, \hat{p}(\theta))$
$\gamma = \int_a^b f(\theta) \, \mathrm{d}\theta$	$\tilde{\gamma} = \frac{b-a}{N_\theta}\sum_{i=1}^{N_\theta} \hat{f}(\theta_i), \quad \theta_1, \ldots, \theta_{N_\theta} \sim \mathrm{U}(a, b)$

TABLE 1. Nested Monte Carlo estimation notation.

For simplicity we consider $[a, b] = [0, 1]$, and we use the union bound to simplify the concentration bound of our problem:

$$\Pr\left(|\hat{\gamma} - \gamma| \geq t\right) = \Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} \hat{f}(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)]\Big| \geq t\right)$$

$$= \Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} \max_{x\in\mathcal{K}_{i,\hat{y}}} \hat{g}(\theta_i, x, \hat{y}) - \mathop{\mathbb{E}}_{\theta}\left[\max_x g(\theta, x, y)\right]\Big| \geq t\right)$$

$$
\begin{aligned}
&= \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i,x,\hat{y}) - \max_{x\in\mathcal{K}_{i,y}} g(\theta_i,x,y)\Big) \\
&\qquad + \frac{1}{N_\theta}\sum_{i=1}^{N_\theta} f(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)]\Big| \geq t\Bigg) \\
&\leq \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} f(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)] \\
&\qquad + \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i,x,\hat{y}) - \max_{x\in\mathcal{K}_{i,y}} g(\theta_i,x,y)\Big)\Big| \geq t\Bigg) \\
&\leq \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} f(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)] \\
&\qquad + \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i,x,\hat{y}) - \max_{x\in\mathcal{K}_{\theta,\hat{y}}} g(\theta,x,\hat{y}) \\
&\qquad + \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_{x\in\mathcal{K}_{\theta,\hat{y}}} g(\theta,x,\hat{y}) - \max_{x\in\mathcal{K}_{\theta,y}} g(\theta,x,y))\Big| \geq t\Bigg) \\
&\leq \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} f(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)]\Big| \geq \frac{t}{3}\Bigg) && (56) \\
&\qquad + \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i,x,\hat{y}) - \max_{x\in\mathcal{K}_{\theta,\hat{y}}} g(\theta,x,\hat{y})\Big| \geq \frac{t}{3}\Bigg) && (57) \\
&\qquad + \Pr\Bigg(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_{x\in\mathcal{K}_{\theta,\hat{y}}} g(\theta,x,\hat{y}) - \max_{x\in\mathcal{K}_{\theta,y}} g(\theta,x,y))\Big| \geq \frac{t}{3}\Bigg) && (58)
\end{aligned}
$$

So we proceed to bound Equations (56) to (58) individually.

Equation (58): Assuming $\max_{z,\theta} p_{\mathcal{M}}(\theta \mid z) = M$, hence $|p(\theta)| \leq M$ and using that g is Lipchitz with respect to y, we obtain

$$
\begin{aligned}
\Pr\Big(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_x g(\theta_i,x,\hat{y}) - \max_x g(\theta_i,x,y)\Big| \geq \frac{t}{3}\Big) &\leq \Pr\Big(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\kappa^+|\hat{p}(\theta_i) - p(\theta_i)| \geq \frac{t}{3}\Big) \\
&\leq \Pr\Big(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}|\hat{p}(\theta_i) - p(\theta_i)| \geq \frac{t}{3\kappa^+}\Big)
\end{aligned}
$$

Given $Z_i = |\hat{p}(\theta_i) - p(\theta_i)|$, since $\hat{p}(\theta_i)$ is the Monte Carlo approximation of $o(\theta)$, we have

$$
\mathbb{E}[Z_i] \leq \sqrt{\mathbb{E}(Z_i^2)} = \sqrt{\mathrm{Var}(\hat{y})} \leq \frac{1}{2\sqrt{N_p}}
$$

where last inequality follows from Popoviciu's inequality applied to $p(\theta_i)$ probabilities, hence bounded by one. Finally, applying Hoeffding's inequality for Z_i we obtain:

$$
\Pr\Big(\sum_{i=1}^{N_\theta} Z_i - \mathbb{E}[Z_i] \geq t\Big) \leq \exp\Big(-\frac{2t^2}{N_\theta}\Big) \tag{59}
$$

$$
\Pr\Big(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} Z_i - \frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\mathbb{E}[Z_i] \geq t\Big) \leq \exp\Big(-2N_\theta t^2\Big) \tag{60}
$$

$$\Pr\Big(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} Z_i \geq t + \mathbb{E}[Z]\Big) \leq \exp\Big(-2N_\theta t^2\Big) \tag{61}$$

$$\Pr\Big(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} Z_i \geq t\Big) \leq \exp\Big(-2N_\theta(t-\mathbb{E}[Z])^2\Big) \leq \exp\Big(-2N_\theta(t-\frac{1}{2\sqrt{N_p}})^2\Big). \tag{62}$$

Summarizing,

$$\Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(g(\theta_i, x^i_{\hat{y}}, \hat{y}) - g(\theta_i, x^i_y, y)\Big)\Big| \geq \frac{t}{3}\right) \leq \mathrm{e}^{-2N_\theta\left(\frac{t}{3\kappa^+} - \frac{1}{2\sqrt{N_p}}\right)^2}$$

Equation (57): Note that,

$$\Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i, x, \hat{y}) - \max_{x\in\mathcal{K}_{i,\hat{y}}} g(\theta_i, x, \hat{y})\Big)\Big| \geq \frac{t}{3}\right) \tag{63}$$

$$\leq \Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\max_{x\in\mathcal{K}_{i,\hat{y}}}\Big(\hat{g}(\theta_i, x, \hat{y}) - g(\theta_i, x, \hat{y})\Big)\Big| \geq \frac{t}{3}\right) \tag{64}$$

$$\leq \Pr\left(\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big|\max_{x\in\mathcal{K}_{i,\hat{y}}}\Big(\hat{g}(\theta_i, x, \hat{y}) - g(\theta_i, x, \hat{y})\Big)\Big| \geq \frac{t}{3}\right) \tag{65}$$

If $k = 1$, then the maximum is completely determined, hence we proceed analogously that for previous case, since $g \in [-M, M]$, $\mathbb{E}[|\hat{g} - g|] \leq \frac{M}{\sqrt{N_z}}$:

$$\Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i, x, \hat{y}) - \max_{x\in\mathcal{K}_{i,\hat{y}}} g(\theta_i, x, \hat{y})\Big)\Big| \geq \frac{t}{3}\right) \leq \mathrm{e}^{-\frac{N_\theta}{2M^2}\left(\frac{t}{3} - \frac{M}{\sqrt{N_z}}\right)^2}. \tag{66}$$

However, if $k \geq 2$, we need to consider the maximum. Note that $X_i = \hat{g} - g$ is a sub-Gaussian variable with $\sigma^2 = \frac{M^2}{N_z}$. Hence, $\max_{i\in\mathcal{K}}|X_i|$ satisfies:

$$\mathbb{E}[\max_{1\leq i\leq k}|X_i|] \leq \sqrt{2\sigma^2\ln(2k)} \leq \frac{M}{\sqrt{N_z}}\sqrt{2\ln(2k)}$$

so applying Hoeffding,

$$\Pr\left(\frac{1}{N_\theta}\Big(\sum_i \max_x |X^i_x| - \sqrt{2\sigma^2\ln 2k}\Big) \geq t\right) \leq \mathrm{e}^{-\frac{N_\theta t^2}{2M^2}} \tag{67}$$

$$\Leftrightarrow \Pr\left(\frac{1}{N_\theta}\sum_i \max_x |X^i_x| \geq t\right) \leq \mathrm{e}^{-\frac{N_\theta}{2M^2}\left(t - \frac{M}{\sqrt{N_z}}\sqrt{2\ln(2k)}\right)^2}. \tag{68}$$

Hence,

$$\Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta}\Big(\max_{x\in\mathcal{K}_{i,\hat{y}}}\hat{g}(\theta_i, x, \hat{y}) - \max_{x\in\mathcal{K}_{i,\hat{y}}} g(\theta_i, x, \hat{y})\Big)\Big| \geq \frac{t}{3}\right) \leq \mathrm{e}^{-\frac{N_\theta}{2M^2}\left(\frac{t}{3} - \frac{M}{\sqrt{N_z}}\sqrt{2\ln(2k)}\right)^2}$$

Equation (56) Follows directly from Hoeffding's inequality with $f \in [-M, M]$:

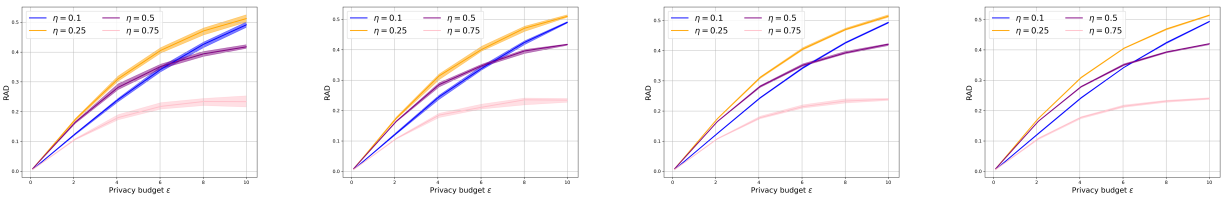
$$\Pr\left(\Big|\frac{1}{N_\theta}\sum_{i=1}^{N_\theta} f(\theta_i) - \mathop{\mathbb{E}}_{\theta}[f(\theta)]\Big| \geq \frac{t}{3}\right) \leq 2\mathrm{e}^{-\frac{N_\theta}{2M^2}\frac{t^2}{9}} \tag{69}$$

Putting all together we have the final error control bound:

$$\Pr(|\hat{\gamma} - \gamma| \geq t) \leq 2\mathrm{e}^{-\frac{N_\theta}{2M^2}\frac{t^2}{9}} + \mathrm{e}^{-\frac{N_\theta}{2M^2}(\frac{t}{3} - \frac{M}{\sqrt{N_z}}\sqrt{2\ln(2k)})^2} + \mathrm{e}^{-\frac{N_\theta}{2M^2}(\frac{t}{3\kappa^+} - \frac{1}{2\sqrt{N_p}})^2}.$$

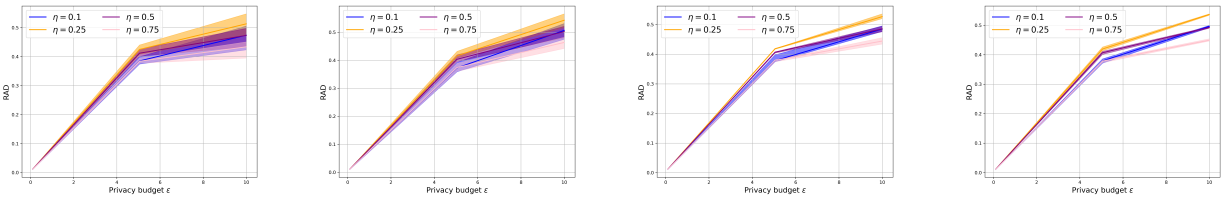
with $k \leq 4$. We empirically evaluate this method for the exponential mechanism [23] with $u = -|z - \theta|$ in $\mathcal{Z} = [0, 1]$, hence,

$$M \leq \frac{1}{s(1 - \mathrm{e}^{-\frac{1}{s}})} \quad \text{where } s = \frac{2\Delta}{\varepsilon},$$



(A) $N = 500$. (B) $N = 1000$. (C) $N = 2000$. (D) $N = 5000$.

FIGURE 3. Numerical approximation of Theorem 4.2 with empirical 95% confidence intervals for the exponential mechanism and $\pi = U(0, 1)$ continuous and $N_p = N_\theta = N_z \equiv N$.



(A) $N = 500$. (B) $N = 1000$. (C) $N = 2000$. (D) $N = 5000$.

FIGURE 4. Numerical approximation of Theorem 4.2 with empirical 95% confidence for the exponential mechanism and $\pi = \text{Beta}(0.1, 0.1)$ continuous and $N_p = N_\theta = N_z \equiv N$.

ensuring the convergence of the numerical estimation. In particular, we test uniform and beta distribution and $\eta \in \{0.1, 0.25, 0.5, 0.75\}$ and report the estimate and the empirical confidence intervals with $J = 500$ repetitions since this ensure a convergence of the intervals for a tolerance of $\tau = 1^{-3}$ for all tested ε. We present the results in Figures 3 and 4, showing the estimated risk along with empirical 95% confidence intervals. The figures illustrate a consistent pattern: the size of the confidence intervals decreases as the sample size increases. As expected, the variance grows with increasing ε, since the corresponding constant $\mathcal{M}$ also increases with ε.

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

Our bound in Theorem 4.2 is universally tight, but two limitations remain. First, it requires full knowledge of the mechanism, making it suitable for noise calibration; however, in DP auditing, we often have only query access (e.g., auditing external software) without insight into the internal protocol [30]. Second, the bound lacks a closed form hence may rely on numerical approximation, particularly for continuous data domains. Consequently, in this section we provide black-box bounds for the case $aux = \varnothing$, both because this is the standard assumption in prior DP auditing [6, 52] and data reconstruction studies [9, 33], and because it makes practical sense: for other auxiliary-information models, one can always rely on the closed-form bound provided by Theorem 4.1.

First, we present a general bound that applies 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

$$\kappa^-_{\pi,\ell}(\eta) = \inf_{z_0 \in \mathcal{Z}} \Pr_{Z \sim \pi} \left[\ell(Z, z_0) \leq \eta\right], \tag{70}$$

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

Theorem 5.1. *If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies f-DP, then any for any attack with $aux = \{\varnothing\}$, $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, it 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. [44][p.23] that for any $S \subseteq \Theta$, for any f-DP mechanism, and $z_0, z_1 \in \mathcal{Z}$,

$$\Pr_{\mathcal{M}}(S \mid D_{z_1}) \leq 1 - f(\Pr_{\mathcal{M}}(S \mid D_{z_0}). \tag{71}$$

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

$$f(\mathbb{E}_X[X]) \leq \mathbb{E}_X[f(X)] \Rightarrow -\mathbb{E}_X[f(X)] \leq f(\mathbb{E}_X[X]). \tag{72}$$

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,

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

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

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

where last inequality follows from Equation (72). 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,

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

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

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

$$\leq (1-\kappa_\pi)\left(1 - f(\sum_{z_1}\sum_{z_0\neq z_1}\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]\tfrac{\pi_0\pi_1}{(1-\kappa_\pi)}) - \sum_{z_1}\sum_{z_0\neq z_1}\Pr[\mathcal{A}_{z_1}(D_{z_0})\in S_\eta(z_1)]\tfrac{\pi_0\pi_1}{(1-\kappa_\pi)}.\right)$$

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 $aux = \{\varnothing\}$. In the following example we see its practical application to Gaussian DP:

Example 7. We consider uniform prior and $\eta = 0$, hence $\kappa^+ = \frac{1}{m}$. Applying Theorem 5.1 we obtain

$$\text{0-RAD} \leq \max_{\alpha\in\left[0,\frac{1}{m-1}\right]} f(\alpha) \;=\; \max_{\alpha\in\left[0,\frac{1}{m-1}\right]} 1-\Phi\Big(\Phi^{-1}(1-\alpha)-\mu\Big)-\alpha,$$

where Φ and φ denote respectively the CDF and PDF of the standard normal distribution.

Using the chain rule and the identity

$$\frac{d}{d\alpha}\Phi^{-1}(1-\alpha) = -\frac{1}{\varphi(\Phi^{-1}(1-\alpha))},$$

we obtain

$$\begin{aligned} f'(\alpha) &= -\varphi\Big(\Phi^{-1}(1-\alpha)-\mu\Big)\cdot\frac{d}{d\alpha}\Big[\Phi^{-1}(1-\alpha)-\mu\Big]-1 \\ &= \frac{\varphi(\Phi^{-1}(1-\alpha)-\mu)}{\varphi(\Phi^{-1}(1-\alpha))}-1. \end{aligned}$$

Moreover, the derivative can be rewritten in closed form. Recall that the standard normal density is

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\,e^{-x^2/2}.$$

Therefore,

$$f'(\alpha) = \frac{\varphi(\Phi^{-1}(1-\alpha)-\mu)}{\varphi(\Phi^{-1}(1-\alpha))}-1 \tag{73}$$

$$= \frac{\exp\Big(-\frac{1}{2}(\Phi^{-1}(1-\alpha)-\mu)^2\Big)}{\exp\Big(-\frac{1}{2}(\Phi^{-1}(1-\alpha))^2\Big)}-1 \tag{74}$$

$$= \exp\left(\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}\right)-1. \tag{75}$$

An interior maximizer satisfies $f'(\alpha)=0$, i.e.,

$$\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}=0$$

Because $\mu>0$, the unique solution is

$$\Phi^{-1}(1-\alpha)=\frac{\mu}{2} \Leftrightarrow \alpha = 1-\Phi(\frac{\mu}{2})$$

Moreover, since

$$f'(\alpha) = \exp\left(\mu\,\Phi^{-1}(1-\alpha)-\frac{\mu^2}{2}\right)-1,$$

we have $f'(\alpha)>0$ for $\alpha<1-\Phi(\mu/2)$ and $f'(\alpha)<0$ for $\alpha>1-\Phi(\mu/2)$. Hence, f increases up to $\alpha^\star$ and decreases thereafter, and the maximizer is unique.

It follows that the unconstrained maximizer is

$$\alpha^\star_{\text{free}} = 1 - \Phi\left(\frac{\mu}{2}\right)$$

Imposing the constraint $\alpha \leq \frac{1}{m-1}$ yields

$$\alpha^\star = \min\left\{\frac{1}{m-1},\ 1 - \Phi\left(\frac{\mu}{2}\right)\right\}.$$

Consequently,

$$0\text{-RAD} \leq \frac{m-1}{m}\left(1 - \Phi\left(\Phi^{-1}(1-\alpha^\star) - \mu\right) - \alpha^\star\right).$$

Moreover, as a consequence of the previous result, we obtain a general result for any (ε, δ)-DP mechanism:

Proposition 5.1. *If a mechanism $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ satisfies (ε, δ)-DP, then for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$, it satisfies*

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

Proof. Follows from combining previous theorem with [21] 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

$$f(\alpha) = \max\{\underbrace{1 - \delta - \mathrm{e}^\varepsilon\alpha}_{f_1(\alpha)}, \underbrace{\frac{1-\delta-\alpha}{\mathrm{e}^\varepsilon}}_{f_2(\alpha)}\}. \tag{76}$$

On the other side, applying Theorem 5.1 we have

$$\eta\text{-RAD} \leq \max_{\alpha\in[\kappa^-,\kappa^+]} (1 - f(\alpha) - \alpha). \tag{77}$$

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:

$$1 - f_1(\alpha) - \alpha = \delta + \mathrm{e}^\varepsilon\alpha - \alpha \tag{78}$$

$$= \alpha(\mathrm{e}^\varepsilon - 1) + \delta \leq \kappa^+(\mathrm{e}^\varepsilon - 1) + \delta \tag{79}$$

Second, for f_2 we obtain:

$$1 - f_2(\alpha) - \alpha = 1 - \frac{1-\delta-\alpha}{\mathrm{e}^\varepsilon} - \alpha \tag{80}$$

$$= 1 - \frac{1-\delta}{\mathrm{e}^\varepsilon} + \alpha(\mathrm{e}^{-\varepsilon} - 1) \leq 1 - \kappa^-(1 - \mathrm{e}^{-\varepsilon}) - \frac{1-\delta}{\mathrm{e}^\varepsilon} = \frac{(1-\kappa^-)(\mathrm{e}^\varepsilon - 1) + \delta}{\mathrm{e}^\varepsilon}. \tag{81}$$

Combined with the general bound Theorem 4.1 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 and do not trivially support partial reconstruction, e.g. [27, 28]. For such settings, we derive more precise bounds. To do so, we first introduce the following auxiliary lemma:

Lemma 5.1. *Given $|\mathcal{Z}| = m$ and $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$ and $\gamma_z = \Pr_{\mathcal{M}}(\Theta_z \mid z) - \Pr_{\mathcal{M}}(\Theta_z)$, with Θ_z as in 23, then*

$$\Gamma := \sum_{z\in\mathcal{Z}} \gamma_z \le \frac{(m-1)(\mathrm{e}^{\varepsilon} - 1 + \delta m)}{\mathrm{e}^{\varepsilon} + m - 1}. \tag{82}$$

Proof. By definition $\Theta_z \cap \Theta_{z'} = \varnothing$. Besides, since for all θ it exits at least one $z_\theta \in \arg\max_z p_{\mathcal{M}}(\theta \mid z)\pi_z$, $\cup\Theta_z = \Theta$. 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

$$\sum_{z\in\mathcal{Z}} \Pr_{\mathcal{M}}(\Theta_z \mid z_0) = \sum_z \int_{\Theta_z} p_{\mathcal{M}}(\theta \mid z_0)\,\mathrm{d}\mu(\theta) = \int_{\Theta} p_{\mathcal{M}}(\theta \mid z_0)\,\mathrm{d}\mu(\theta) = 1. \tag{83}$$

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

$$\Pr_{\mathcal{M}}(\Theta_1 \mid z_0) \ge \mathrm{e}^{-\varepsilon}(\Pr_{\mathcal{M}}(\Theta_1 \mid z_1) - \delta). \tag{84}$$

substituting Equation (84) in Equation (83) we obtain, for all $i, j \in [m]$,

$$\Pr_{\mathcal{M}}(\Theta_i \mid z_i) + \mathrm{e}^{-\varepsilon} \sum_{i\neq j} \Pr_{\mathcal{M}}(\Theta_j \mid z_j) \le 1 + \delta \mathrm{e}^{-\varepsilon}(m-1) \tag{85}$$

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

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) + (m-1)\mathrm{e}^{-\varepsilon} \sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) \le m(1 + \delta \mathrm{e}^{-\varepsilon}(m-1)) \Leftrightarrow \tag{86}$$

$$\sum_{i=1}^{m} \Pr_{\mathcal{M}}(\Theta_i \mid z_i) \le \frac{m(1 + \delta \mathrm{e}^{-\varepsilon}(m-1))}{1 + (m-1)\mathrm{e}^{-\varepsilon}} = \frac{m\mathrm{e}^{\varepsilon} + \delta m(m-1)}{\mathrm{e}^{\varepsilon} + (m-1)}. \tag{87}$$

Hence,

$$\Gamma = \sum_{z\in\mathcal{Z}} \gamma_z \tag{88}$$

$$= \sum_{z\in\mathcal{Z}} \left(\Pr_{\mathcal{M}}(\Theta_z \mid z) - \Pr_{\mathcal{M}}(\theta_z) \right) \tag{89}$$

$$= \sum_{z\in\mathcal{Z}} \Pr_{\mathcal{M}}(\Theta_z \mid z) - 1 \tag{90}$$

$$\le \frac{m\mathrm{e}^{\varepsilon} + \delta m(m-1)}{\mathrm{e}^{\varepsilon} + m - 1} - 1 = \frac{(m-1)(\mathrm{e}^{\varepsilon} - 1 + \delta m)}{\mathrm{e}^{\varepsilon} + m - 1}. \tag{91}$$

□

Applying this lemma we obtain the following RAD bound:

Theorem 5.2 (0-RAD under (ε, δ)-DP)**.** *Given $|\mathcal{Z}| = m$ with prior $\pi_1(1-\pi_1) \ge \cdots \ge \cdots \ge \pi_m(1-\pi_m)$ and $\mathcal{M}\colon \mathcal{Z}^n \to \mathcal{D}(\Theta)$ an (ε, δ)-DP mechanism, for any attack $A\colon \Theta \to \mathcal{D}(\mathcal{Z})$*

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

where $K \in [m]$ is the largest index satisfying $R = (m-1)\frac{e^{\varepsilon}-1+m\delta}{e^{\varepsilon}+m-1} - (K - \sum_{i=1}^{K}\pi_i)\frac{e^{\varepsilon}-1+2\delta}{e^{\varepsilon}+1} \ge 0$ and $K_\pi = \sum_i^K (1-\pi_i)\pi_i$.

Proof. Since $|\mathcal{Z}| = m$ and $aux = \varnothing$, Theorem 4.2 gets reduced to *Equation* (23), hence

$$0\text{-RAD} \leq \sum_{i=1}^{m} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) \right) \pi_i \equiv \sum_{i=1}^{m} \gamma_i \, \pi_i. \tag{92}$$

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

$$\gamma_i = \Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i) = \int_{\Theta_i} p_{\mathcal{M}}(\theta \mid z_i) - \sum_{j \in [m]} p_{\mathcal{M}}(\theta \mid z_j) \pi_j \, \mathrm{d}\mu(\theta) \tag{93}$$

$$= \int_{\Theta_i} \sum_{j \in [m]} \left(p_{\mathcal{M}}(\theta \mid z_i) - p_{\mathcal{M}}(\theta \mid z_j) \right) \pi_j \, \mathrm{d}\mu(\theta) \tag{94}$$

$$= \sum_{j \neq i} \left(\Pr_{\mathcal{M}}(\Theta_i \mid z_i) - \Pr_{\mathcal{M}}(\Theta_i \mid z_j) \right) \pi_j \tag{95}$$

$$\leq \mathrm{TV}(\mathcal{M}) \sum_{j \neq i} \pi_j \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} (1 - \pi_i). \tag{96}$$

If we simply apply this bound we recover Theorem 4.1 result:

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

However, due to Lemma 5.1, we know that this bound is loose, since in this case,

$$\Gamma = \sum_{i=1}^{m} \gamma_i = \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} (m - 1) \geq \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1} (m - 1) = \Gamma_{\max}, \tag{97}$$

contradicting Lemma 5.1; 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,

$$\sum_{r=1}^{k} \gamma_{i_r} \leq \frac{e^{\varepsilon} - 1 + m\delta}{e^{\varepsilon} + m - 1} (m - 1) \Leftrightarrow \tag{98}$$

$$\sum_{r=1}^{k} (1 - \pi_{i_r}) \leq (m - 1) \frac{(\mathrm{e}^{\varepsilon} - 1 + m\delta)((\mathrm{e}^{\varepsilon} + 1)}{(\mathrm{e}^{\varepsilon} - 1 + 2\delta)((\mathrm{e}^{\varepsilon} - 1 + 2\delta))} \tag{99}$$

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,

$$0\text{-RAD} \leq \frac{e^{\varepsilon} - 1 + 2\delta}{e^{\varepsilon} + 1} \sum_{i=1}^{k_\pi} \pi_i(1 - \pi_i) + R \max_{r > k_\pi} \pi_r \tag{100}$$

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 4.1. This formulation enables the assessment of intermediate configurations of π. Notably, when $\pi = U[m]$ yields a marked improvement:

Corollary 5.1 (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 $A = \frac{e^\varepsilon - 1 + 2\delta}{e^\varepsilon + 1}$ and applying Theorem 5.2 we get:

$$0\text{-RAD} \leq AK\frac{m-1}{m^2} + \frac{1}{m}(\Gamma - K\frac{m-1}{m}A) = \frac{1}{m}\Gamma = \frac{e^\varepsilon - 1 + \delta m}{e^\varepsilon + m - 1}\frac{m-1}{m}. \tag{101}$$

□

Remark on Composition. Since our η-RAD bounds depend explicitly on the privacy parameters–namely ε, δ, and/or f–they can be directly recomputed under composition by first applying the corresponding composition results to obtain the composed privacy parameters (Cf.Section 2), and then evaluating the bounds on these composed values. In the following example, we illustrate how to derive RAD composition bounds for the particular case of DP-SGD.

Example 8. Given a risk threshold, RAD $\leq \gamma$, we aim to calibrate the noise scale σ on a full-batch DP-SGD (i.e., the standard deviation of the Gaussian noise added to the gradients during training [1]), for T steps to protect against the threat model considered by Hayes, Balle, and Mahloujifar [33], i.e., white-box access to private gradients, uniform prior over $|\mathcal{Z}| = m$ and $\eta = 0$, hence $\kappa_- = \kappa_+ = 1/m$.

Each iteration of a full-batch DP-SGD is (σ^{-1})-GDP [21], hence by f-DP composition rule, T iterations of DP-SGD are $(\sqrt{T}\sigma^{-1})$-GDP (cf. Section 2.2). Combining this composition result with our theorems we obtain direct calibration rules:

Without information about *aux*, we use Theorem 4.1. Any μ-GDP mechanism has total variation $\mathrm{TV} \leq 2\Phi(\frac{\mu}{2}) - 1$ [29], hence DP-SGD after T iterations satisfies $\gamma \leq \frac{m-1}{m}(\Phi(\frac{\sqrt{T}\sigma^{-1}}{2}) - 1)$. We plot this bound for $T = 100$ in Figures 5b and 5c.

If we consider the whole records sensitive, $aux = \{\varnothing\}$, then we apply Theorem 5.1:

$$0\text{-RAD} \leq \frac{m-1}{m}\max_{\alpha \in [0, \frac{1}{m-1}]}\left(1 - \Phi\left(\Phi^{-1}(1-\alpha) - \frac{\sqrt{T}}{\sigma}\right) - \alpha\right)$$

Hence, given $\alpha^* = \min\left\{\frac{1}{m-1},\ 1 - \Phi\left(\frac{\sqrt{T}}{2\sigma}\right)\right\}$, the minimum σ to guarantee 0-RAD $\leq \gamma$ is:

$$\sigma \geq \frac{\sqrt{T}}{\Phi^{-1}(1-\alpha^*) - \Phi^{-1}\left(1 - \frac{m}{m-1}\gamma - \alpha^*\right)}.$$

We plot this bound for the case of $T = 100$ in Figure 5a.

In summary, this section provides reasonable closed-form upper bounds (as we show in Section 7.3) for estimating RAD when Theorem 4.2 cannot be computed explicitly or $\mathcal{M}$ is unknown and $aux = \{\varnothing\}$, hence Theorem 4.1 would overestimate the risk. Importantly, these bounds offer composition results as we summarize in Table 2.

Notion	Assumptions	RAD bound	Composition	ReRo bound
Total variation	—	Theorem 4.1	✓	∄
f-DP	$aux = \{\varnothing\}$	Theorem 5.1	✓	Equation (3) [33]
$\mathcal{M}$	aux known	Theorem 4.2	×	∄
(ε, δ)-DP	—	Theorem 4.1	✓	∄
(ε, δ)-DP	$aux = \{\varnothing\}$	Proposition 5.1	✓	Equation (1) [9]
(ε, δ)-DP	$aux = \{\varnothing\}, \eta = 0$	Theorem 5.2	✓	Equation (1) [9]

TABLE 2. Summary of RAD bounds applicability.

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 [4, 13, 38]. 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 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 $\tilde{\gamma}$, we invert our theoretical bounds 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, in previous sections, we provide bounding functions B such that $\mathrm{RAD}(\mathcal{M}) \leq B(\varepsilon, \delta)$ for any (ε, δ)-DP mechanism. 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 completely black-box scenario–where not even the mechanism used is known–for categorical data, in which we assume $\pi = U[m]$, the best bound is Corollary 5.1. Therefore, the DP auditing framework consists of running an attack, measuring its empirical RAD $\widetilde{\gamma}$, and deriving $\widetilde{\varepsilon}$ as follows:

$$\widetilde{\varepsilon} = \begin{cases} \ln\left(\frac{\widetilde{\gamma}\frac{m}{m+1}+1}{1-\widetilde{\gamma}\frac{m}{m-1}}\right) & \text{if the term can be evaluated,} \\ \text{undefined} & \text{otherwise.} \end{cases} \tag{102}$$

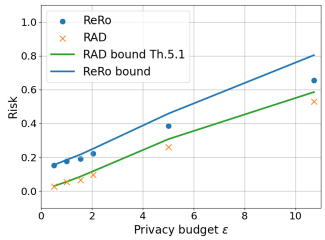
However, if the mechanism $\mathcal{M}$ is known, we can use our improved bound from Theorem 4.2 (See examples 1 to 3).

Our auditing framework overcomes the fundamental scalability limitations of prior learning-based approaches such as DP-Sniper and Eureka [13, 50], enabling auditing in high-dimensional categorical LDP settings. Unlike these methods, our approach avoids costly hyperparameter tuning and the search for worst-case neighboring databases, and remains computationally feasible even when the input domain contains thousands of categories (see Section 7).

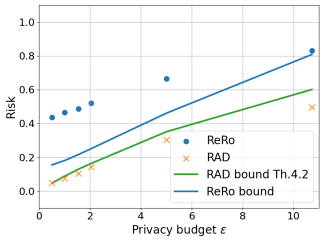
Despite the importance of LDP mechanisms [25], only one major work has so far focused on LDP auditing: LDP Auditor [6]. Applying our RAD-based DP auditing to LDP, we address key limitations of prior work. In contrast to LDP Auditor, which focuses exclusively on perfect reconstruction without target-specific 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) [6], 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 8 for results), where we audit three main LDP mechanisms–GRR, SS and OUE–showing improved accuracy for all of them.

7. Experiments

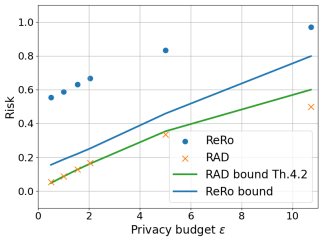
In this section, we empirically examine the limitations of ReRo described in Section 3, focusing on how existing bounds fail to account for realistic attackers with target-specific auxiliary information. Moreover, we validate our theoretical bounds and our RAD-based DP auditing framework in real-world 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 [6].



(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 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.

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

For DP-SGD, we use the same dataset as in ReRo [33] for consistency: MNIST [46], with 70,000 grayscale images of handwritten digits. We also replicate results on Fashion-MNIST [70] (Fashion), which similarly contains 70,000 grayscale images of clothing items.

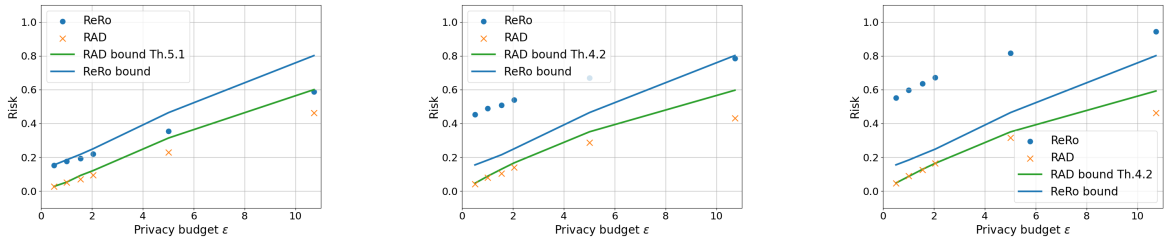
To evaluate the imputation attack [41], we use the Census and Texas-100X datasets in consistency with the original paper. The Census dataset [41] contains 1,676,013 records with 14 attributes, where race is treated as the sensitive attribute with eight categories. The Texas-100X dataset [41] comprises 925,128 patient records from 441 hospitals, including demographic and medical attributes, with a binary ethnicity attribute designated to be sensitive.

We evaluate aggregation in the Adult dataset [11], a census dataset commonly used in privacy-preserving aggregation [64]. It consists of $32,561$, records with two numerical attributes, from which we select (working) hours-per-week following previous work [64], leading to the domain $\mathcal{Z} = \{0, \ldots, 100\}$.

We evaluate our LDP auditing framework on location-reconstruction attacks using two real-world mobility datasets: the Porto dataset [60] and the Geolife dataset [74]. Both datasets are widely used in privacy and mobility research (e.g., [48, 62, 71]) and are publicly available. Each dataset consists of GPS coordinates, which we map to the OpenStreetMap (OSM) graph format [61] 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 investigate attacks on private learning (DP-SGD), aggregation queries (Laplace mechanism), and LDP protocols (GRR, OUE, SS) under varying auxiliary information settings to validate our bounds, compare RAD and ReRo, and evaluate our auditing framework.

We demonstrate *ReRo overestimating risk* due to imputation and how RAD overcomes this with the pure imputation attack [41]: It 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 [41], we tested in both the Census and Texas datasets. We set $|D_-| = 49,000$



(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 6. 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.

and a universe $\mathcal{Z}$ of $m = 1000$, randomly selected from the remaining data records consistent with [41]. We define the attack to be successful, $\ell(z, z') = 0$, if $a(z) = a(z')$, as is typical for AIAs.

We show *RAD improvement over ReRo* and the optimality of our bounds both in private learning and DP aggregation. In both cases, we test our optimal attacks to assess tightness.

For private learning we run the attacks against DP-SDG on the MNIST and Fashion image datasets in three settings: $aux = z$ (a MIA), $aux = \{\varnothing\}$ (a DRA, replicating the setting in [33]), and $aux = a(z)$ (a DRA, where the adversary also knows the target image's label, i.e., which object is contained). 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 the cling rate, i.e., the maximum norm we clip the real gradients to while training, $C = 0.1$ and $\delta = 10^{-5}$ and adjust the noise scale σ (see Example 8) 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 [33] parameters.

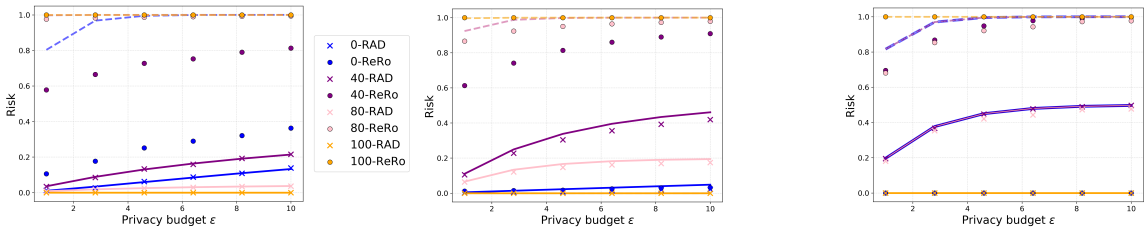
For DP aggregation, we evaluate the optimal attack against the Laplace mechanism on sum queries using the "working-hours" attribute of Adult, employing truncation as a post-processing operation. We empirically compute the distribution π from the original data to simulate a real-world setting (reflecting that working 40 h/week is apriori more likely than working 100 h/week), a uniform distribution, and a completely skewed distribution with $\pi(100) = \pi(0) = 1/2$. For all cases, we set $|D| = 999$, $aux = \{\varnothing\}$ and evaluate the performance for $\eta \in \{0, 40, 80, 100\}$.

Finally, *we evaluate RAD in LDP*, and we compare our auditing framework with the state-of-the-art tool LDP AUDITOR [6] for three relevant LDP mechanisms: GRR, OUE and SS [7, 32]. To obtain the results for LDP AUDITOR, we used the code from Arcolezi and Gambs's public GitHub repository [5]. LDP AUDITOR estimates the empirical privacy budget in 10^6 runs.

We evaluate RAD based on our optimal attack (See Alg. 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 4.2 and obtain an estimate of the empirical privacy budget. The precise $B(\varepsilon)$ for GRR, OUE and SS are shown in Examples 1 to 3 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 [63]. We set $\tau = 10^{-6}$. Consistent with [6], we repeat the ε estimation five times and report the mean and standard deviation.

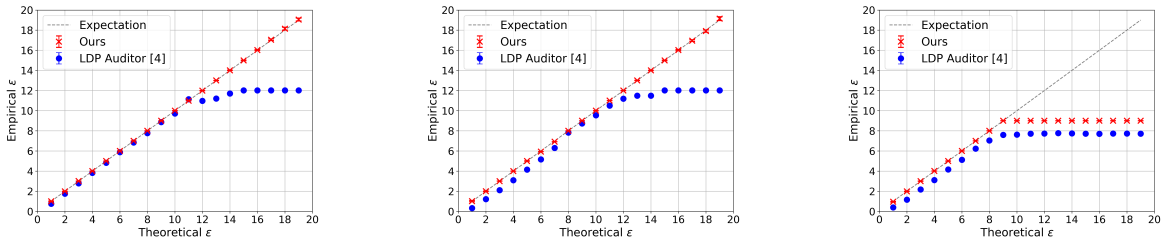
All experiments use empirical estimates of ReRo and RAD. Following [33], ReRo is estimated by repeating the attack $A(\mathcal{M}(D_z), a(z))$ J times for each $z \in \mathcal{Z}$ and computing the π-weighted average. The RAD correction term is estimated analogously by evaluating $A(\mathcal{M}(D_{z_0}), a(z_1))$ J times for each target–challenger pair $z_1, z_0 \in \mathcal{Z}$ and averaging the results.

For MNIST, Fashion and Adult, we set $J = 1,000$ (as in [33]). 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



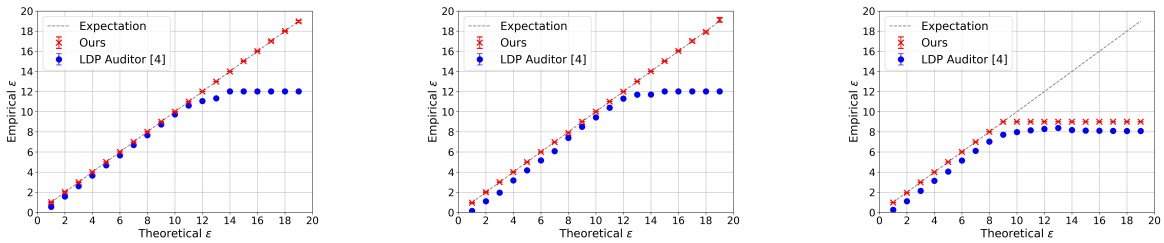
(A) Adult distribution. (B) Uniform distribution. (C) $\pi(0) = \pi(100) = \frac{1}{2}$.

FIGURE 7. RAD vs ReRo results for optimal attack against Truncated Laplace on Adult. Straight lines show RAD bounds (Theorem 4.2) and dashed lines ReRo bounds ([33]). Markers show empirical risk as estimated by RAD/ReRo.



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

FIGURE 8. LDP Audit results from RAD-based auditing and LDP AUDITOR [6] on Porto dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.



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

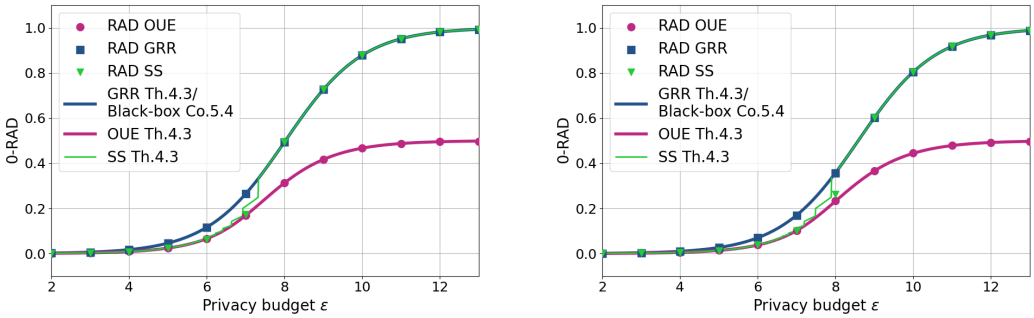
FIGURE 9. LDP Audit results from RAD-based auditing and LDP AUDITOR [6] on Geolife dataset. Values along the diagonal indicate perfect accuracy; below it, privacy is overestimated; above it, underestimated.

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 TensorFlow [67] to evaluate the attacks. For DP-SGD ReRo, we rely on a minimal implementation provided by Hayes, Balle, and Mahloujifar [33], which we extend to incorporate RAD and target-specific auxiliary knowledge. For the imputation attack [41], we adapt the authors' public implementation [40].

7.3. **Results.** In this section, we present the results of RAD and ReRo empirical risk estimates and their corresponding theoretical bounds. For both ReRo and RAD, the y-axis shows the risk measure, with values near one indicating high risk and near zero indicating low risk.

7.3.1. *RAD covers, but ReRo breaks for auxiliary knowledge.* Figure 5 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 6. We also include the



(A) Porto (B) Geolife

FIGURE 10. RAD results for LDP mechanisms. Lines show theoretical bounds and markers empirical RAD.

Dataset	ReRo	RAD
Census	0.81	0
Texas	0.73	0

TABLE 3. ReRo Vs. RAD risk estimation for imputation attack.

corresponding theoretical bounds for ReRo and RAD for comparison. As expected, the existing bounds for ReRo [33] correctly upper-limit the empirically observed ReRo risk when the adversary has no prior knowledge of the victim record ($aux = \{\varnothing\}$, Figure 5a). However, when the adversary has prior knowledge of the victim record (Figures 5b and 5c), 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 5b and 5c show that the tightness of our worst-case bound Theorem 4.1 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 5a shows that our closed-form bound Theorem 5.1 offers a reasonable upper-bound also when Theorem 4.2 needs to be numerically approximated (as is the case, for instance, with DP-SGD).

7.3.2. *Leakage vs. Imputation.* Table 3 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.

This tendency of ReRo to overestimate risk is not confined to this setting. In our optimal attacks on DP-SGD (Figure 5), ReRo consistently overestimates leakage across all investigated cases, with the effect becoming more pronounced as more auxiliary information is incorporated. Membership inference ($a(z) = z$) provides the clearest example, where ReRo reports risk values exceeding 0.6 even for privacy budgets $\varepsilon \leq 4$, which are commonly considered to offer strong privacy guarantees [47]. This behavior aligns with expectations, as ReRo cannot discount auxiliary information; consequently, greater attacker knowledge leads to larger overestimation.

Similarly, Figure 7 shows that ReRo fails to capture the effect of the success threshold η. As η increases, an oblivious attacker's success probability rises, but ReRo cannot account for

this since it depends only on success probability and thus converges to 1 for all ε. This results in substantial overestimation: for $\eta = 100$, a trivial setting where any guess is correct, ReRo reports maximal risk despite the mechanism providing no advantage. In contrast, RAD properly discounts this effect, showing that increasing η boosts advantage only up to a point (here, $\eta = 40$), after which the advantage decreases as success becomes nearly granted.

7.3.3. *Bound tightness.* Figure 7 shows the results of RAD and ReRo for our optimal attack against Laplace mechanism on Adult including their corresponding theoretical bounds. Figures 10a and 10b shows the analogous for LDP mechanisms, GRR, OUE and SS, on the Porto and Geolife datasets. On the x-axis, we see ε and the y-axes the exact estimated risk for such ε selection. Note that for LDP, RAD and ReRo results coincide, since the attack relies solely on the released output (with no auxiliary information or imputation effects). Moreover, the prior-based chance level under the uniform prior is negligible for $|\mathcal{Z}| = 3,052$. We therefore report only RAD to avoid redundancy.

We observe that our bounds (cf. Theorem 4.2) are tight for every prior π and capture even subtle differences between mechanisms. In particular, the RAD estimates for GRR perfectly match our perfect-reconstruction black-box bound (Theorem 5.2), confirming its tightness.

Moreover, Figure 7 clearly illustrates the impact of the data distribution: the skewed distribution (Figure 7c) constitutes the worst case, while the empirical distribution represents the best case. This highlights that knowledge of the data distribution can substantially improve utility; in the absence of such knowledge, the only safe choice is calibration with respect to the worst case.

Finally, these results provide concrete evidence for the importance of attack-based noise calibration. For identical values of ε, OUE offers significantly stronger protection against DRAs than GRR and SS. Hence, ε alone does not capture the full privacy picture, and RAD is essential for understanding the actual privacy implications of a mechanism for users.

7.3.4. *Auditing Local DP with RAD.* Figures 8 and 9 shows the results from our LDP Auditing experiments using the Porto and Geolife datasets. 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 [6]. In contrast, the tightness of our RAD bound enables our auditing approach to accurately estimate empirical privacy budgets for the whole range, without such a 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 [6]: as we prove in Example 2, 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 4.2 enables precise and reliable auditing based on DRAs.

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 [9], drastically overestimates the actual leakage of DP mechanisms when target-specific public knowledge exists–leading to excessive utility loss if used as noise calibration methods. 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 private learning, DP aggregation and LDP 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 et al. "Deep Learning with Differential Privacy". In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. CCS '16. New York, USA: ACM, 2016, pp. 308–318. DOI: 10.1145/2976749.2978318.

[2] J. M. Abowd. "The U.S. Census Bureau Adopts Differential Privacy". In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '18. London, United Kingdom: Association for Computing Machinery, 2018, p. 2867. ISBN: 9781450355520. DOI: 10.1145/3219819.3226070.

[3] M. S. M. S. Annamalai and E. De Cristofaro. "Nearly tight black-box auditing of differentially private machine learning". In: *Proceedings of the 38th International Conference on Neural Information Processing Systems*. NIPS '24. Vancouver, BC, Canada: Curran Associates Inc., 2025. ISBN: 9798331314385.

[4] M. S. M. S. Annamalai et al. *The Hitchhiker's Guide to Efficient, End-to-End, and Tight DP Auditing*. 2025. arXiv: 2506.16666.

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

[6] 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 (2024), pp. 123–141. DOI: 10.56553/popets-2024-0110.

[7] H. H. Arcolezi et al. "On the Risks of Collecting Multidimensional Data Under Local Differential Privacy". In: *Proceedings of the VLDB Endowment* 16.5 (2023), pp. 1126–1139. DOI: 10.14778/3579075.3579086.

[8] F. Baccelli, B. Blaszczyszyn, and M. K. Karray. *Random Measures, Point Processes, and Stochastic Geometry*. hal-02460214: Inria, July 2024. URL: https://inria.hal.science/hal-02460214.

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

[10] 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*. Vol. 80. Stockholm, Sweden: PMLR, 2018, pp. 394–403. URL: https://proceedings.mlr.press/v80/balle18a.html.

[11] B. Becker and R. Kohavi. *Adult*. UCI Machine Learning Repository. 1996. DOI: doi.org/10.24432/C5XW20.

[12] D. Bernau et al. "Quantifying identifiability to choose and audit ϵ in differentially private deep learning". In: *Proceedings of the VLDB Endowment* 14.13 (2021), pp. 3335–3347. DOI: 0.14778/3484224.348423.

[13] B. Bichsel et al. “DP-Sniper: Black-Box Discovery of Differential Privacy Violations using Classifiers”. In: *Symposium on Security and Privacy (SP)*. San Francisco, USA: IEEE, 2021, pp. 391–409. DOI: 10.1109/SP40001.2021.00081.

[14] Y. Bu et al. “Privacy preserving serial data publishing by role composition”. In: *Proceedings of the VLDB Endowment* 1.1 (2008), pp. 845–856.

[15] M. Bun et al. *Statistical inference is not a privacy violation*. https://differentialprivacy.org/inference-is-not-a-privacy-violation/. Accessed: 2025-06-10. 2021.

[16] C. Carey et al. “Measuring re-identification risk”. In: *Proceedings of the ACM on Management of Data* 1.2 (2023), pp. 1–26.

[17] K. Chatzikokolakis et al. “Bayes security: A not so average metric”. In: *36th Computer Security Foundations Symposium (CSF)*. Dubrovnik, Croatia: IEEE, 2023, pp. 388–406. DOI: 10.1109/CSF57540.2023.00011.

[18] G. Cormode et al. “Synthetic Tabular Data: Methods, Attacks and Defenses”. In: *Proceedings of the VLDB Endowment* 18.12 (2025), pp. 5448–5450. DOI: 10.14778/3750601.3750692.

[19] T. Cunningham et al. “Real-world trajectory sharing with local differential privacy”. In: *Proceedings of the VLDB Endowment* 14.11 (2021), pp. 2283–2295.

[20] Z. Ding et al. “Detecting Violations of Differential Privacy”. In: *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*. CCS’18. New York, USA: ACM, 2018, pp. 475–489.

[21] J. Dong, A. Roth, and W. J. Su. “Gaussian Differential Privacy”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 84.1 (2022), pp. 3–37. DOI: 10.1111/rssb.12454.

[22] C. Dwork. “Differential Privacy”. In: *Automata, Languages and Programming*. Berlin, Heidelberg: Springer, 2006, pp. 1–12. DOI: 10.1007/11787006_1.

[23] 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. DOI: 10.1561/0400000042.

[24] C. Dwork et al. “Calibrating noise to sensitivity in private data analysis”. In: *Theory of Cryptography: Third Theory of Cryptography Conference*. New York, USA: Springer, 2006, pp. 265–284. DOI: 10.1007/11681878_14.

[25] Ú. Erlingsson, V. Pihur, and A. Korolova. “RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response”. In: *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’14. New York, USA: ACM, 2014, pp. 1054–1067. DOI: 10.1145/2660267.2660348.

[26] Ú. Erlingsson et al. *That which we call private*. 2019. arXiv: 1908.03566.

[27] 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. New York, USA: ACM, 2015, pp. 1322–1333. DOI: 10.1145/2810103.2813677.

[28] M. Fredrikson et al. “Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing”. In: *23rd USENIX Security Symposium*. San Diego, CA: USENIX Association, 2014, pp. 17–32.

[29] E. Ghazi and I. Issa. “Total variation meets differential privacy”. In: *IEEE Journal on Selected Areas in Information Theory* 5 (2024), pp. 207–220. DOI: 10.1109/JSAIT.2024.3384083.

[30] D. Gorla et al. “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.

[31] 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, USA: ACM, 2024, pp. 155–171. DOI: 10.1145/3689943.3695046.

[32] M. E. Gursoy et al. "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. DOI: 10.1109/TIFS.2022.3170242.

[33] J. Hayes, B. Balle, and S. Mahloujifar. "Bounding training data reconstruction in DP-SGD". In: *Advances in Neural Information Processing Systems*. Vol. 36. New Orleans, USA: Curran Associates, Inc., 2023, pp. 78696–78722.

[34] X. He, N. Raval, and A. Machanavajjhala. "A demonstration of VisDPT: Visual exploration of differentially private trajectories". In: *Proceedings of the VLDB Endowment* 9.13 (2016), pp. 1489–1492.

[35] W. Hoeffding. "Probability inequalities for sums of bounded random variables". In: *Journal of the American statistical association* 58.301 (1963), pp. 13–30.

[36] F. Houssiau et al. *TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data*. 2022. arXiv: 2211.06550.

[37] T. Humphries et al. "Investigating Membership Inference Attacks under Data Dependencies". In: *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*. Los Alamitos, CA, USA: IEEE, 2023, pp. 473–488. DOI: 10.1109/CSF57540.2023.00013.

[38] 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.

[39] 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.

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

[41] 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.

[42] 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*. ICML'16. New York, USA: JMLR.org, 2016, pp. 2436–2444.

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

[44] D. Kifer et al. *Bayesian and frequentist semantics for common variations of differential privacy: Applications to the 2020 census*. 2022. eprint: 209.03310.

[45] B. Kulynych et al. "Attack-aware noise calibration for differential privacy". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 134868–134901. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/f33e853ba1f5f038268f9839e37821d5-Paper-Conference.pdf.

[46] Y. LeCun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[47] 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.

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

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

[50] Y. Lu et al. "Eureka: A General Framework for Black-box Differential Privacy Estimators". In: *Symposium on Security and Privacy (SP)*. San Francisco, USA: IEEE, 2024, pp. 913–931. DOI: 10.1109/SP54263.2024.00166.

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

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

[53] M. Malek et al. "Antipodes of label differential privacy: PATE and ALIBI". In: *Proceedings of the 35th International Conference on Neural Information Processing Systems*. NIPS '21. Red Hook, NY, USA: Curran Associates Inc., 2021. ISBN: 9781713845393.

[54] F. D. McSherry. "Privacy integrated queries: an extensible platform for privacy-preserving data analysis". In: *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*. SIGMOD '09. Providence, Rhode Island, USA: Association for Computing Machinery, 2009, pp. 19–30. ISBN: 9781605585512. DOI: 10.1145/1559845.1559850.

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

[56] Y.-A. de Montjoye et al. "Unique in the Crowd: The privacy bounds of human mobility". In: *Nature Scientific Reports* 3 (2013). DOI: 10.1038/srep01376.

[57] P. Nanayakkara et al. "What are the chances? explaining the epsilon parameter in differential privacy". In: *Proceedings of the 32nd USENIX Conference on Security Symposium*. SEC '23. USA: USENIX Association, 2023. ISBN: 978-1-939133-37-3.

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

[59] M. Nasr et al. "Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning". In: *Symposium on Security and Privacy (SP)*. San Francisco, USA: IEEE, 2021, pp. 866–882.

[60] 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.

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

[62] 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.

[63] T. Sauer. *Numerical Analysis*. 2nd. USA: Addison-Wesley Publishing Company, 2011. ISBN: 0321783670.

[64] J. Soria-Comas et al. "Enhancing data utility in differential privacy via microaggregation-based k-anonymity". In: *The VLDB Journal* 23.5 (2014), pp. 771–794.

[65] 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.

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

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

[68] F. Tramèr et al. *Debugging Differential Privacy: A Case Study for Privacy Auditing*. 2022. eprint: 2202.12219.

[69] T. Wang et al. "Locally differentially private protocols for frequency estimation". In: *26th USENIX Security Symposium*. USA: USENIX Association, 2017, pp. 729–745.

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

[71] 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: ACM, 2015, pp. 1298–1309. ISBN: 9781450338325. DOI: 10.1145/2810103.2813640.

[72] 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. 759–763. DOI: 10.1109/ISIT.2017.8006630.

[73] S. Yeom et al. “ Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting ”. In: *31st Computer Security Foundations Symposium (CSF)*. Los Alamitos, CA, USA: IEEE, 2018, pp. 268–282. DOI: 10.1109/CSF.2018.00027.

[74] Y. Zheng et al. *Geolife GPS trajectory dataset - User Guide*. July 2011. URL: https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/.