Residual-PAC Privacy: Automatic Privacy Control Beyond the Gaussian Barrier

Tao Zhang
Washington University in St. Louis

Yevgeniy Vorobeychik
Washington University in St. Louis

Abstract

The Probably Approximately Correct (PAC) Privacy framework [56] provides a powerful instance-based methodology to preserve privacy in complex data-driven systems. Existing PAC Privacy algorithms (we call them Auto-PAC) rely on a Gaussian mutual information upper bound. However, we show that the upper bound obtained by Auto-PAC is tight if and only if under the data distribution, the unperturbed output is Gaussian and the noise is independent Gaussian. We propose two approaches for addressing this issue. First, we introduce two tractable post-processing methods for Auto-PAC, based on Donsker–Varadhan representation and sliced Wasserstein distances. However, the result still leaves "wasted" privacy budget. To address this issue more fundamentally, we introduce Residual-PAC (R-PAC) Privacy, an f-divergence-based measure to quantify privacy that remains after adversarial inference. To implement R-PAC Privacy in practice, we propose a Stackelberg Residual-PAC (SR-PAC) automatic privatization algorithm, a game-theoretic framework that selects optimal noise distributions through convex bilevel optimization. Our approach achieves efficient privacy budget utilization for arbitrary data distributions and naturally composes when multiple mechanisms access the dataset. Our experiments demonstrate that SR-PAC obtains consistently a better privacy-utility tradeoff than both PAC and differential privacy baselines.

1 Introduction

Data-driven decision systems power critical applications ranging from medical diagnosis to autonomous vehicles, yet their outputs can inadvertently expose sensitive information contained in the data. As data pipelines grow in scale and complexity, practitioners need rigorous and scalable privacy guarantees that go beyond ad-hoc testing. Over the past two decades, formal privacy frameworks have proliferated. Differential Privacy (DP) [17] (and its variants such as Rényi DP [40]) delivers input-independent worst-case indistinguishability by bounding output distribution shifts from single-record changes. Alternative information-theoretic definitions, such as mutual-information DP [13], Fisher-information bounds [20, 26, 28], and Maximal Leakage [31, 48], provide complementary guarantees and offer alternative trade-offs between privacy and utility.

Nevertheless, provable privacy guarantees for modern data-processing algorithms remains a challenge. First, worst-case frameworks like DP require computing global sensitivity, which is generally NP-hard [58]. Moreover, computing the optimal privacy bound of DP under composition is, in general, #P-complete [41]. In practice, finding the minimal noise needed to meet a target guarantee is intractable for most real-world algorithms, especially when the effect of each operation on privacy is unclear. On the other hand, empirical or simulation-based methods (e.g., testing resistance to membership inference [51]) address specific threats but lack rigorous, adversary-agnostic assurance. Bridging this gap requires a new, broadly applicable framework that can quantify and enforce privacy risk without relying on sensitivity.

A promising alternative has recently emerged: the Probably Approximately Correct (PAC) Privacy framework [56]. PAC Privacy shifts from indistinguishability-based guarantees to an operational notion that measures the *information-theoretic hardness* of reconstructing sensitive information. It is defined by an impossibility-of-inference guarantee for a chosen adversarial task and data prior, and the framework provides algorithms that enforce tractable mutual-information upper bounds to certify this guarantee. This approach enables automatic privatization via black-box simulation, and enjoys additive composition bounds and automatic privacy budget implementations for adaptive sequential compositions of mechanisms with arbitrary interdependencies. Notably, PAC Privacy often requires only $O(1)$ noise magnitude to achieve its privacy guarantees (independent of the output dimension), whereas differential privacy's worst-case, input-independent noise magnitude scales as $\Theta(\sqrt{d})$ for a d-dimensional release.

However, existing PAC privacy algorithms (which we refer to as Auto-PAC) are fundamentally conservative. In particular, we show (Proposition 1) that Auto-PAC achieves the

designated privacy budget exactly if and only if under the data distribution, the unperturbed output is Gaussian and the noise is independent Gaussian, so that the unperturbed and perturbed outputs are jointly Gaussian. Consequently, Auto-PAC will in general make inefficient use of the privacy budget. Conservative privacy accounting is a central practical concern in DP and PAC Privacy because conservative bounds impose unnecessary noise and waste privacy budget, particularly under composition. Narrowing this conservativeness remains an open challenge in PAC privacy [57].

We address this limitation of Auto-PAC in two ways. First, working within the general PAC Privacy framework, we develop two tractable post-processing methods for Auto-PAC conservativeness reduction, based on Donsker–Varadhan representation [15] and sliced Wasserstein distances [5,47]. However, even these methods fail to fully close the privacy budget gap. To address this issue more fundamentally, we introduce the notion of *Residual-PAC Privacy* (R-PAC privacy). Unlike PAC privacy, which aims to quantify and bound the privacy leaked, R-PAC privacy focuses instead on quantifying *privacy remaining after information has been leaked by a data processing mechanism*, using f-divergence to this end. When f-divergence is instantiated as Kullback–Leibler (KL) divergence, we show that Residual-PAC Privacy is fully characterized by the conditional entropy up to a known constant that does not depend on the mechanism or the applied noise.

To implement R-PAC privacy with KL divergence, we propose a novel *Stackelberg Residual-PAC (SR-PAC)* automatic privatization algorithm. SR-PAC formulates the problem of privatization via noise perturbation, given a privacy budget, as a Stackelberg game in which the leader selects a noise distribution with the goal of minimizing the magnitude of the perturbation, while the follower chooses a stochastic inference strategy to recover the sensitive data. We show that when the entire probability space is considered, the resulting bilevel optimization problem becomes a convex program. Moreover, we prove that the mixed-strategy Stackelberg equilibrium of this game yields the optimal noise distribution, ensuring that the conditional entropy of the perturbed mechanism precisely attains the specified privacy budget. Finally, our experimental evaluation demonstrates that the proposed SR-PAC privacy framework consistently outperforms both PAC-privacy and differential privacy baselines.

A complete appendix, including all proofs, is provided in the online extended version of this paper [60]. We summarize our main contributions as follows:

- We characterize the conservativeness of Auto-PAC [53, 56], showing that it arises from the gap between the surrogate Gaussian mutual information bound and the true non-Gaussian mutual information of the privatized mechanism.
- We propose two computationally tractable approaches to reduce this gap: one based on the Donsker-Varadham representation (Theorem 3) and the other based on the sliced Wasserstein distances (Theorem 4).
- We propose a novel privacy framework, Residual-PAC (R-PAC), to quantify the portion of privacy that remains rather than the amount leaked. This offers a complementary perspective to PAC privacy, and enables efficient implementation of tight privacy budget.
- We present an automatic privatization algorithm, Stackelberg R-PAC (SR-PAC), to efficiently compute noise distributions for a given privacy budget. SR-PAC algorithm achieves tight budget utilization, can operate with only black-box access via Monte Carlo simulation, and adaptively concentrates noise in privacy-sensitive directions while preserving task-relevant information.

1.1 Related Work

Privacy Quantification Notions. Differential privacy (DP) and its variants have become the gold standard for formal privacy quantification and guarantees, with the original definitions by Dwork et al. [17,18] formalizing privacy loss through bounds on the distinguishability of outputs under neighboring datasets. Variants such as concentrated differential privacy (CDP) [8, 19], zero-concentrated DP (zCDP) [7], and Rényi differential privacy (RDP) [40] have further extended this framework by parameterizing privacy loss with different statistical divergences (e.g., Rényi divergence), thereby enhancing flexibility in privacy accounting, especially for compositions and adaptive mechanisms. Pufferfish privacy [34,46,52,59,61] generalizes DP by considering secrets that go beyond DP's presence and absence of individual records. Information-theoretic measures provide alternative and complementary approaches for quantifying privacy loss. For instance, mutual information has been used to analyze privacy leakage in a variety of settings [10, 13], with f-divergence and Fisher information offering finer-grained or context-specific metrics [20, 26, 28, 56]. These frameworks help to bridge the gap between statistical risk and adversarial inference, and are closely connected to privacy-utility trade-offs in mechanism design. Maximal Leakage [31], hypothesis testing interpretations [3], and other relaxations further broaden the analytic toolkit for measuring privacy risk.

Privacy-Utility Trade-off. Balancing the trade-off between privacy and utility is a central challenge in the design of privacy-preserving mechanisms. This challenge is frequently formulated as an optimization problem [2, 16, 22, 23, 25, 27, 36, 39, 49]. For example, Ghosh et al. [23] demonstrated that the geometric mechanism is universally optimal for DP under certain loss-minimizing criteria in Bayesian settings, while Lebanon et al. [36] and Alghamdi et al. [2] studied utility-constrained optimization. Gupte et al. [27] modeled the privacy-utility trade-off as a zero-sum game between privacy mechanism designers and adversaries, illustrating the

interplay between optimal privacy protection and worst-case adversarial loss minimization.

Optimization Approaches for Privacy. A growing body of work frames the design of privacy-preserving mechanisms as explicit optimization problems, aiming to maximize data utility subject to formal privacy constraints. Many adversarial or game-theoretic approaches—such as generative adversarial privacy (GAP) [29] and related GAN-based frameworks [11, 32, 42]—cast the privacy mechanism designer and the adversary as players in a min-max game, optimizing utility loss and privacy leakage, respectively. More recently, Selvi et al. [50] introduced a rigorous optimization framework for DP based on distributionally robust optimization (DRO), formulating the mechanism design problem as an infinite-dimensional DRO to derive noise-adding mechanisms that are nonasymptotically and unconditionally optimal for a given privacy level. Their approach yields implementable mechanisms via tractable finite-dimensional relaxations, often outperforming classical Laplace or Gaussian mechanisms on benchmark tasks. Collectively, these lines of research illustrate the power of optimization and game-theoretic perspectives in achieving privacy-utility trade-offs beyond conventional privatization mechanisms.

2 Preliminaries

2.1 PAC Privacy

Privacy Threat Model. We consider the following general privacy problem. A sensitive input X (e.g., a dataset, membership status) is drawn from a distribution $\mathcal{D}$, which may be unknown or inaccessible. There is a data processing (possibly randomized) mechanism $\mathcal{M}: \mathcal{X} \mapsto \mathcal{Y} \subset \mathbb{R}^d$, where $\mathcal{Y}$ is measurable. An adversary observes the output $Y = \mathcal{M}(X)$ and attempts to estimate the original input X with an estimation $\widetilde{X}$. The adversary has complete knowledge of both the data distribution $\mathcal{D}$ and the mechanism $\mathcal{M}$, representing the worst-case scenario. The central privacy concern is determining whether the adversary can accurately estimate the true input, meeting some predefined success criterion captured by an indicator function ρ. The PAC privacy framework [56] addresses this threat model and is formally defined as follows.

Definition 1 ($(\delta, \rho, \mathcal{D})$ PAC Privacy [56])**.** *For a data processing mechanism $\mathcal{M}$, given some data distribution $\mathcal{D}$, and a measure function $\rho(\cdot,\cdot)$, we say $\mathcal{M}$ satisfies* $(\delta, \rho, \mathcal{D})$ PAC Privacy *if the following experiment is impossible:*

A user generates data X from distribution $\mathcal{D}$ and sends $\mathcal{M}(X)$ to an adversary. The adversary who knows $\mathcal{D}$ and $\mathcal{M}$ is asked to return an estimation $\widetilde{X} \in \mathcal{X}$ on X such that with probability at least $1-\delta$, $\rho(\widetilde{X}, X) = 1$.

Definition 1 formalizes privacy in terms of the adversary's difficulty in achieving accurate reconstruction, capturing the semantics of the *impossibility of customized adversarial inference* [57]. The function ρ specifies the success criterion for reconstruction, adapting to the requirements of the specific application. For example, when $\mathcal{X} \subset \mathbb{R}^{d'}$, one may define success as $\rho(\widetilde{X}, X) \equiv \mathbf{1}\{|\widetilde{X} - X|_2 \leq \varepsilon\} = 1$ or some small $\varepsilon > 0$ with $\mathbf{1}\{\cdot\}$ as the indicator. If X is a finite set of size n, success may be defined as correctly recovering more than $n - \varepsilon$ elements. Notably, ρ need not admit a closed-form expression; it simply indicates whether the reconstruction satisfies the designated criterion for success.

In PAC Privacy, X may represent general *secrets* as considered in Pufferfish privacy frameworks [34, 46, 52, 59, 61], which go beyond data points. For example, a secret may be a dataset attribute or a global feature of the dataset. For ease of exposition, this paper focuses on the setting where X denotes the data. PAC Privacy treats the secrets X as a random variable drawn from a distribution $\mathcal{D}$. When $\mathcal{D}$ is not available in closed form, we may explicitly create $\mathcal{D}$ via a sampling rule and access it through i.i.d. samples from a data pool [57].

PAC Privacy considers the following adversarial worst-case scenario: a computationally unbounded adversary with full knowledge of both $\mathcal{D}$ and the underlying query function. The randomness inherent in data (or secret) generation and the randomness in the query function are the only elements unknown to the adversary [56, 57]. PAC Privacy is highly flexible by enabling ρ to encode a wide range of adversary models and user-specified risk criteria. For example, in membership inference attacks [9], $\rho(\widetilde{X}, X) = 1$ may indicate that $\widetilde{X}$ successfully determines the presence of a target data point in X. In reconstruction attacks [4], success may be defined by $\rho(\widetilde{X}, X) = 1$ if $|\widetilde{X} - X|_2 \leq 1$, representing a close approximation of the original data.

Given the data distribution $\mathcal{D}$ and the adversary's criterion ρ, the *optimal prior success rate* $(1 - \delta_o^\rho)$ is defined as the highest achievable success probability for the adversary without observing the output $\mathcal{M}(X)$: $\delta_o^\rho = \inf_{\widetilde{X}_0} \Pr_{X \sim \mathcal{D}}\left(\rho(\widetilde{X}_0, X) \neq 1\right)$. Similarly, the *posterior success rate* $(1-\delta)$ is defined as the adversary's probability of success after observing $\mathcal{M}(X)$.

The notion of *PAC advantage privacy* quantifies how much the mechanism output $\mathcal{M}(X)$ can improve the adversary's success rate, based on f-divergence.

Definition 2 (f-Divergence)**.** *Given a convex function $f: (0, +\infty) \to \mathbb{R}$ with $f(1) = 0$, extend f to $t = 0$ by setting $f(0) = \lim_{t \to 0^+} f(t)$ (in $\mathbb{R} \cup \{+\infty, -\infty\}$). The f-divergence between two probability distributions P and Q over a common measurable space is:*

$$\mathtt{D}_f(P\|Q) \equiv \begin{cases} \mathbb{E}_Q\left[f\left(\frac{dP}{dQ}\right)\right] & \text{if } P \ll Q, \\ +\infty & \text{otherwise,} \end{cases}$$

where $\frac{dP}{dQ}$ is the Radon-Nikodym derivative.

Definition 3 ($(\Delta_f^\delta, \rho, \mathcal{D})$ PAC Advantage Privacy [56])**.** *A mechanism $\mathcal{M}$ is termed $(\Delta_f^\delta, \rho, \mathcal{D})$* PAC advantage private *if it is $(\delta, \rho, \mathcal{D})$ PAC private and*

$$\Delta_f^\delta \equiv \mathcal{D}_f(\mathbf{1}_\delta \| \mathbf{1}_{\delta_o^\rho}) = \delta_o^\rho f(\frac{\delta}{\delta_o^\rho}) + (1-\delta_o^\rho) f(\frac{1-\delta}{1-\delta_o^\rho}).$$

Here, $\mathbf{1}_\delta$ and $\mathbf{1}_{\delta_o^\rho}$ represent two Bernoulli distributions of parameters δ and δ_o^ρ, respectively.

Here, PAC Advantage Privacy is defined on top of PAC Privacy and quantifies the amount of *privacy loss* incurred from releasing $\mathcal{M}(X)$, captured by the additional *posterior advantage* Δ_f^δ.

2.2 Automatic PAC Privatization Algorithms

PAC Privacy enables automatic privatization, which supports simulation-based implementation for arbitrary black-box mechanisms, without requiring the worst-case adversarial analysis, such as sensitivity computation. In this section, we present the main theorems and algorithms underlying automatic PAC privatization as introduced in [56] (hereafter "Auto-PAC") and the efficiency-improved version proposed in [53] (hereafter "Efficient-PAC"; algorithm details in Appendix D in [60]). We start by defining the *mutual information*.

Definition 4 (Mutual Information)**.** *For random variables A and B, the mutual information is defined as*

$$\mathtt{MI}(A;B) \equiv \mathcal{D}_{KL}(\mathsf{P}_{A,B} \| \mathsf{P}_A \otimes \mathsf{P}_B),$$

the KL-divergence between their joint distribution (i.e., $\mathsf{P}_{A,B}$) and the product of their marginals (i.e., P_A and P_B).

When the f-divergence in Δ_f^δ is instantiated as the KL divergence (denoted as $\Delta_{\mathrm{KL}}^\delta$), Theorem 1 of [56] shows

$$\Delta_{\mathrm{KL}}^\delta \le \mathtt{MI}(X; \mathcal{M}(X)). \tag{1}$$

That is, we can control the posterior advantage $\Delta_{\mathrm{KL}}^\delta$ by bounding the mutual information between private data and the released output. Importantly, this mutual information bound holds uniformly over all adversarial inference procedures (including the choice of ρ) permitted by PAC Privacy. Therefore, when $\Delta_f^\delta = \Delta_{\mathrm{KL}}^\delta$, we can characterize and quantify the PAC Privacy in terms of $\mathtt{MI}(X; \mathcal{M}(X))$ without requiring any adversarial model or ρ tuning, while the semantics of PAC Privacy remains as the impossibility of customized adversarial inference.

Next, we introduce the Auto-PAC. Consider a deterministic mechanism $\mathcal{M}: \mathcal{X} \to \mathbb{R}^d$, where the output norm is uniformly bounded: $\|\mathcal{M}(X)\|_2 \le r$ for all X. To guarantee PAC Privacy, the mechanism is perturbed by Gaussian noise $B \sim \mathcal{N}(0, \Sigma_B)$, where Σ_B is the covariance. When $X \sim \mathcal{D}$, let $\Sigma_{\mathcal{M}(X)}$ be the covariance of $\mathcal{M}(X)$. For any deterministic mechanism $\mathcal{M}$ and any Gaussian noise B, define the *Gaussian surrogate bound*

$$\mathtt{LogDet}(\mathcal{M}(X), B) \equiv \frac{1}{2} \log\det\left(I_d + \Sigma_{\mathcal{M}(X)} \cdot \Sigma_B^{-1}\right). \tag{2}$$

Algorithm 1 $(1-\gamma)$-Confidence Auto-PAC [56]

Require: deterministic mechanism $\mathcal{M}$, dataset $\mathcal{D}$, sample size m, security parameter c, mutual information quantities β' and v.

for $k = 1, 2, \ldots, m$ **do**
  Generate $X^{(k)}$ from $\mathcal{D}$. Record $y^{(k)} = \mathcal{M}(X^{(k)})$.
end for
Calculate $\hat{\mu} = \sum_{k=1}^m y^{(k)}/m$ and $\hat{\Sigma} = \sum_{k=1}^m (y^{(k)} - \hat{\mu})(y^{(k)} - \hat{\mu})^\top / m$.
Apply SVD: $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^\top$, where $\hat{\Lambda}$ has eigenvalues $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \ldots \ge \hat{\lambda}_d$.
Find $j_0 = \arg\max_j \hat{\lambda}_j$ for $\hat{\lambda}_j > c$.
if $\min_{1 \le j \le j_0, 1 \le l \le d} |\hat{\lambda}_j - \hat{\lambda}_l| > r\sqrt{dc} + 2c$ **then**
  for $j = 1, 2, \ldots, d$ **do**
    Set $\lambda_{B,j} = \dfrac{2v}{\sqrt{\hat{\lambda}_j + 10cv/\beta'} \cdot \left(\sum_{j=1}^d \sqrt{\hat{\lambda}_j + 10cv/\beta'}\right)}$.
  end for
  Set $\Sigma_{\boldsymbol{B}} = \hat{U}\Lambda_{\boldsymbol{B}}^{-1}\hat{U}^\top$.
else
  Set $\Sigma_{\boldsymbol{B}} = (\sum_{j=1}^d \hat{\lambda}_j + dc)/(2v) \cdot \boldsymbol{I}_d$.
end if
Output: $\Sigma_{\boldsymbol{B}}$.

Theorem 1 (Theorem 3 of [56])**.** *For an arbitrary deterministic mechanism $\mathcal{M}$ and Gaussian noise $B \sim \mathcal{N}(0, \Sigma_B)$, the mutual information satisfies*

$$\mathtt{MI}(X; \mathcal{M}(X) + B) \le \mathtt{LogDet}(\mathcal{M}(X), B).$$

Moreover, there exists Σ_B such that $\mathbb{E}[\|B\|_2^2] = \left(\sum_{j=1}^d \sqrt{\lambda_j}\right)^2$ with $\{\lambda_j\}$ being the eigenvalues of $\Sigma_{\mathcal{M}(X)}$, and $\mathtt{MI}(X; \mathcal{M}(X) + B) \le \frac{1}{2}$.

Theorem 1 establishes a simple upper bound on the mutual information with Gaussian noise perturbation. Choosing Σ_B to implement the Gaussian surrogate bound $\mathtt{LogDet}(\mathcal{M}(X), B) = \beta$ for a privacy budget β enables *anisotropic* noise as it estimates the eigenvectors of $\mathcal{M}(X)$ to fit the noise to the geometry of the eigenspace of $\mathcal{M}(X)$. The result extends naturally to randomized mechanisms (Corollary 2 of [56]). Building on Theorem 1, Algorithm 1 (we refer to it as $(1-\gamma)$-*Confidence Auto-PAC*) is proposed by [56] to perform automatic PAC privatization. Algorithm 1 aims to determine an Gaussian noise covariance Σ_B, so that $\mathtt{MI}(X; \mathcal{M}(X) + B) \le \beta$ is satisfied with confidence at least $1 - \gamma$.

2.3 Differential Privacy

In addition to the standard PAC Privacy, we also compare our approach to the differential privacy (DP) framework. Let $x = (x_1, x_2, \ldots, x_n) \in \mathcal{X} = (\mathcal{X}^\dagger)^n$ be the input dataset, where each data point x_i is defined over some measurable domain $\mathcal{X}^\dagger$. We say two datasets $x, x' \in \mathcal{X}$ are *adjacent* if they differ in exactly one data point.

Definition 5 ($(\varepsilon, \bar{\delta})$-Differential Privacy [18])**.** *A randomized mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ is said to be* $(\varepsilon, \bar{\delta})$-differentially private (DP)*, with $\varepsilon \geq 0$ and $\bar{\delta} \in [0,1]$, if for any pair of adjacent datasets x, x', and any measurable $\mathcal{W} \subseteq \mathcal{Y}$, it holds that* $\Pr[\mathcal{M}(x) \in \mathcal{W}] \leq e^{\varepsilon} \Pr[\mathcal{M}(x') \in \mathcal{W}] + \bar{\delta}$.

The parameter ε is usually referred to as the *privacy budget*, and $\delta \in (0,1]$ represents the failure probability. DP is an input-independent adversarial worst-case approaches that focus on the sensitivity magnitude, while Auto-PAC is instance-based and adds anisotropic noise tailored to each direction as needed. Appendix A characterizes the difference between DP, PAC Privacy, and our Residual-PAC (R-PAC) Privacy.

3 Characterizing The Gaussian Barrier of Automatic PAC Privatization

This section characterizes the utility of Auto-PAC by focusing on the conservativeness of the implemented mutual information bounds. To distinguish from Algorithm 1 ($(1-\gamma)$-confidence Auto-PAC), Auto-PAC refers to the direct implementation of privacy budgets for the bound $\texttt{LogDet}(\mathcal{M}(X), B)$ without a target conference level. The Gaussian surrogate bound is conservative due to a nonzero *Gaussianity gap*, the discrepancy between the *true mutual information* and $\texttt{LogDet}(\mathcal{M}(X), B)$ defined by (2):

$$\texttt{Gap}_{\texttt{d}} \equiv \texttt{LogDet}(\mathcal{M}(X), B) - \texttt{MI}(X; \mathcal{M}(X) + B). \quad (3)$$

Define $Z = \mathcal{M}(X) + B$ with mean $\mu_Z = \mu_{\mathcal{M}(X)}$ and covariance $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$. Let $P_{\mathcal{M},B}$ denote the true distribution of $Z = \mathcal{M}(X) + B$, and define the *Gaussian surrogate distribution* as

$$\widetilde{Q}_{\mathcal{M}} \equiv \mathcal{N}(\mu_Z, \Sigma_Z) \quad (4)$$

with the same first and second moments as $Z \sim P_{\mathcal{M},B}$.

Proposition 1. *Let $B \sim \mathcal{N}(0, \Sigma_B)$. Then, $\texttt{Gap}_{\texttt{d}} = \texttt{D}_{\texttt{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}}) \geq 0$. Moreover, $\texttt{Gap}_{\texttt{d}} = 0$ iff $P_{\mathcal{M},B} = \widetilde{Q}_{\mathcal{M}}$.*

Proposition 1 shows that the conservativeness of $\mathcal{M}(X)$ in terms of the Gaussianity gap is equivalent to the KL divergence between the true output distribution and the Gaussian surrogate distribution. Thus, Auto-PAC tightly implements a privacy budget if and only if the true perturbed output distribution coincides with the Gaussian surrogate distribution in (1).

Proposition 2. *For any privacy budget $\beta > 0$, the noise distribution $Q = \mathcal{N}(0, \Sigma_B)$ obtained by Auto-PAC is the unique solution of the following problem:*

$$\inf_{Q' = \mathcal{N}(\mu, \Sigma')} \mathbb{E}_{B \sim Q'}\left[\|B\|_2^2\right] \quad \text{s.t.} \quad \texttt{MI}(X; \widetilde{Z}) \leq \beta \text{ with } \widetilde{Z} \sim \widetilde{Q}_{\mathcal{M}}. \quad (5)$$

Proposition 2 implies that, if we replace $Z \sim P_{\mathcal{M},B}$ by $\widetilde{Z} \sim \widetilde{Q}_{\mathcal{M}}$, Auto-PAC's zero-mean Gaussian noise is the optimal solution to minimize the magnitude of the Gaussian noise subject to the mutual information constraint.

Proposition 3. *For the same privacy budget $\beta > 0$, let Q and Q_γ, respectively, be the Gaussian noise distribution obtained by Auto-PAC and $(1-\gamma)$-Confidence Auto-PAC with any $\gamma \in [0,1]$. Let $B \sim Q$ and $B_\gamma \sim Q_\gamma$. Then, the following holds.*

(i) $\texttt{MI}(X; \mathcal{M}(X) + B_\gamma) \leq \texttt{MI}(X; \mathcal{M}(X) + B)$.

(ii) $\mathbb{E}_{Q_\gamma}[\|B_\gamma\|_2^2] \geq \mathbb{E}_Q[\|B\|_2^2]$.

In Proposition 3, part (i) shows that $(1-\gamma)$-confidence Auto-PAC is more conservative than directly implementing $\texttt{LogDet}(\mathcal{M}(X), B)$ (Auto-PAC) for the same privacy budget. Part (ii) demonstrates that $(1-\gamma)$-confidence Auto-PAC uses larger noise magnitude than Auto-PAC for the same privacy budget. Thus, in subsequent comparisons involving PAC Privacy, we focus on Auto-PAC.

3.1 Mechanism Comparison in PAC Privacy

Definition 9 of [56] defines the optimal perturbation for PAC Privacy that tightly implements the privacy budget while maintaining optimal utility, where utility is captured by a loss function $\mathcal{K}$. An optimal perturbation Q^* is a solution of the following optimization problem:

$$\inf_{Q} \mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B; \mathcal{M})] \quad \text{s.t.} \quad \texttt{MI}(X; \mathcal{M}(X) + B) \leq \beta, B \sim Q. \quad (6)$$

The choice of utility loss function $\mathcal{K}$ is context-dependent. However, in many applications, we are primarily concerned with the expected Euclidean norm of the noise or a convex function thereof, e.g., $\mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B; \mathcal{M})] = \mathbb{E}_Q\left[\|B\|_2^2\right]$.

We now show in Proposition 4 that using $\mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B; \mathcal{M})] = \mathbb{E}_Q\left[\|B\|_2^2\right]$ is sufficient to obtain perturbations that maintain *coherent ordering* of PAC Privacy using mutual information (i.e., larger privacy budgets yield non-decreasing actual mutual information).

Proposition 4. *Fix a mechanism $\mathcal{M}$ and data distribution $\mathcal{D}$. Let $\mathcal{Q}$ denote the collection of all zero-mean noise distributions under consideration, and let $\texttt{I}_{\text{true}} : \mathcal{Q} \mapsto \mathbb{R}_{\geq 0}$ be the true mutual information functional; i.e., $\texttt{I}_{\text{true}}(Q) = \texttt{MI}(X; \mathcal{M}(X) + B)$ with $B \sim Q$ for $Q \in \mathcal{Q}$. For each privacy budget $\beta \geq 0$, define the feasible region $\mathcal{F}(\beta) \equiv \{Q \in \mathcal{Q} : \texttt{I}_{\text{true}}(Q) \leq \beta\}$. Suppose that $\mathcal{F}(\beta)$ is nonempty for all privacy budgets of interest.*

For each $\beta \geq 0$, let $Q^*(\beta)$ be a solution of the problem:

$$\min_{Q} \mathbb{E}_{B\sim Q}[\|B\|_2^2] \quad s.t. \quad Q \in \mathcal{F}(\beta). \tag{7}$$

Then, if $\beta_1 < \beta_2$, we have $\mathtt{I}_{\text{true}}(Q^*(\beta_1)) \leq \mathtt{I}_{\text{true}}(Q^*(\beta_2))$.

However, if Auto-PAC is used to solve the optimization problem (5), we have the conservative implementation of a given privacy budget. For any mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$, we let $\mathtt{Gap}_{\mathtt{d}}(Q) = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}})$ with $B \sim Q$. The next result shows that when $\mathtt{Gap}_{\mathtt{d}}(Q) > 0$, Auto-PAC does not, in general, maintain coherent ordering of PAC Privacy.

Theorem 2. *Fix a mechanism $\mathcal{M}$ and data distribution $\mathcal{D}$. Let $\mathcal{Q}$ denote the collection of all zero-mean Gaussian distributions under consideration, and let $\mathtt{I}_{\text{true}} : \mathcal{Q} \mapsto \mathbb{R}_{\geq 0}$ be the true mutual information functional; i.e., $\mathtt{I}_{\text{true}}(Q) = \mathtt{MI}(X; \mathcal{M}(X) + B)$ with $B \sim Q$ for $Q \in \mathcal{Q}$. For each $\beta \geq 0$, let $Q^*(\beta)$ be a solution of the optimization in Proposition 2. For any $0 < \beta_1 < \beta_2$, define $\mathtt{G}(\beta_2, \beta_1) \equiv \mathtt{Gap}_{\mathtt{d}}(Q^*(\beta_2)) - \mathtt{Gap}_{\mathtt{d}}(Q^*(\beta_1))$. Then:*

(i) If $\mathtt{G}(\beta_2, \beta_1) \leq \beta_2 - \beta_1$, then $\mathtt{I}_{\text{true}}(Q^*(\beta_1)) \leq \mathtt{I}_{\text{true}}(Q^*(\beta_2))$.

(ii) If $\mathtt{G}(\beta_2, \beta_1) > \beta_2 - \beta_1$, then $\mathtt{I}_{\text{true}}(Q^*(\beta_1)) > \mathtt{I}_{\text{true}}(Q^*(\beta_2))$.

Theorem 2 characterizes when Auto-PAC maintains coherent ordering of actual information leakage $\mathtt{I}_{\text{true}} = \beta - \mathtt{Gap}_{\mathtt{d}}$, and when not. Increasing the budget from β_1 to β_2 permits extra leakage $\beta_2 - \beta_1$ by using Auto-PAC, but part may be wasted if the mechanism output becomes more non-Gaussian. The wasted portion is $\mathtt{G}(\beta_2, \beta_1) = \mathtt{Gap}_{\mathtt{d}}(Q^*(\beta_2)) - \mathtt{Gap}_{\mathtt{d}}(Q^*(\beta_1))$. If this waste exceeds the budget increase, then $\mathtt{I}_{\text{true}}$ decreases despite a larger nominal budget, violating coherent ordering. This result cautions against comparing mechanisms using Auto-PAC solely by budgets, as identical budgets may yield different true PAC Privacy leakages depending on their respective Gaussianity gaps.

3.2 $\mathtt{Gap}_{\mathtt{d}}$ Reduction via Non-Gaussianity Correction

In this section, we propose two approaches to reduce $\mathtt{Gap}_{\mathtt{d}}$ after a $\mathcal{N}(0, \Sigma_B)$ is determined by Auto-PAC. For any deterministic mechanism $\mathcal{M}$ and Gaussian noise $B \sim \mathcal{N}(0, \Sigma_B)$, recall the Gaussian surrogate distribution $\widetilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$ in (4). Let $\mathtt{D}_Z = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}})$. By Proposition 1, $\mathtt{Gap}_{\mathtt{d}} = \mathtt{D}_Z$. For any estimator $\widehat{\mathtt{D}}_Z$ of $\mathtt{D}_Z$, define the *improved mutual information* estimate:

$$\mathtt{IMI}(\widehat{\mathtt{D}}_Z) \equiv \mathtt{LogDet}(\mathcal{M}(X), B) - \widehat{\mathtt{D}}_Z.$$

For $0 \leq \widehat{\mathtt{D}}_Z \leq \mathtt{D}_Z$, we have $\mathtt{MI}(X; \mathcal{M}(X) + B) \leq \mathtt{IMI}(\widehat{\mathtt{D}}_Z) \leq \mathtt{LogDet}(\mathcal{M}(X), B)$. Thus, if we can get $\widehat{\mathtt{D}}_Z$ between $\mathtt{D}_Z$ and 0 after Auto-PAC privatization is performed, then for any Σ_B that ensures $\mathtt{LogDet}(\mathcal{M}(X), B) = \beta$, we have $\mathtt{IMI}(\widehat{\mathtt{D}}_Z) = \beta - \widehat{\mathtt{D}}_Z$ as surrogate upper bound that is tighter than $\mathtt{LogDet}(\mathcal{M}(X), B)$. Thus, we can have tighter privacy accounting post-hoc to the Auto-PAC privatization to save additional privacy budget, without requiring direct mutual information estimation.

Before describing the approaches, we first introduce two standard discrepancy measures between $P_{\mathcal{M},B}$ and $\widetilde{Q}_{\mathcal{M}}$.

Definition 6 (Donsker–Varadhan (DV) Objective [15])**.** *For probability measures P and Q on a common measurable space,*

$$\mathtt{D}_{\mathrm{KL}}(P\|Q) = \sup_{f:\mathcal{Y}\to\mathbb{R}} \left\{ \mathbb{E}_P[f(Y)] - \log \mathbb{E}_Q\left[e^{f(Y)}\right] \right\},$$

where the supremum ranges over measurable f such that $\mathbb{E}_Q[e^f] < \infty$. We call $\mathcal{J}(f; P, Q) \equiv \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]$ the DV objective. In our setting, $\mathtt{D}_Z = \mathtt{D}_{\mathrm{KL}}(P_Z \| \widetilde{Q}_{\mathcal{M}}) = \sup_f \mathcal{J}\left(f; P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}}\right)$.

Definition 7 (Sliced Wasserstein Distances (SWD) [5, 47])**.** *For $p \geq 1$, the p-Wasserstein distance between P and Q on $\mathbb{R}^d$ is $\mathrm{W}_p(P,Q) = \left(\inf_{\eta \in \widehat{\Pi}(P,Q)} \mathbb{E}_{(X,Y)\sim\eta}\left[\|X - Y\|_2^p\right] \right)^{1/p}$, where $\widehat{\Pi}(P,Q)$ is the set of couplings with marginals P and Q. The sliced p-Wasserstein distance averages 1-D Wasserstein distances over directions v on the unit sphere $\mathbb{S}^{d-1}$:*

$$\mathrm{SW}_p^p(P,Q) = \int_{\mathbb{S}^{d-1}} \mathrm{W}_p^p\left(\mathcal{L}(\langle v, X\rangle), \mathcal{L}(\langle v, Y\rangle)\right) d\sigma(v),$$

where σ is the uniform (Haar) measure on $\mathbb{S}^{d-1}$ and $\mathcal{L}(\cdot)$ denotes the law of its argument. In our setting we write $\mathrm{W}_p(P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}})$ and $\mathrm{SW}_p(P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}})$.

Definition 8 (Finite-Sample Lower-Confidence DV Estimator)**.** *Fix a function class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ with $0 \in \mathcal{F}$ and let $\widehat{\mathcal{J}}(f; S_P, S_Q) \equiv \frac{1}{|S_P|}\sum_{z\in S_P} f(z) - \log\left(\frac{1}{|S_Q|}\sum_{z\in S_Q} e^{f(z)}\right)$ denote the empirical DV objective on samples S_P from P_Z and S_Q from $\widetilde{Q}_{\mathcal{M}}$. Draw four independent splits $S_P^{\mathrm{tr}}, S_Q^{\mathrm{tr}}, S_P^{\mathrm{val}}, S_Q^{\mathrm{val}}$ with sizes $n_P^{\mathrm{tr}}, n_Q^{\mathrm{tr}}, n_P^{\mathrm{val}}, n_Q^{\mathrm{val}}$ respectively, and fit $\widehat{f}_{\mathrm{tr}} \in \arg\max_{f\in\mathcal{F}} \widehat{\mathcal{J}}\left(f; S_P^{\mathrm{tr}}, S_Q^{\mathrm{tr}}\right)$.*

Let $\Gamma_{\hat{\delta}} = \Gamma_{\hat{\delta}}\left(\mathcal{F}, n_P^{\mathrm{val}}, n_Q^{\mathrm{val}}\right)$ be any valid uniform deviation bound satisfying, with probability at least $1 - \hat{\delta}$, $\sup_{f\in\mathcal{F}} \left|\widehat{\mathcal{J}}\left(f; S_P^{\mathrm{val}}, S_Q^{\mathrm{val}}\right) - \mathcal{J}\left(f; P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}}\right)\right| \leq \Gamma_{\hat{\delta}}$, where $\mathcal{J}(f; P, Q)$ is the DV objective (Definition 6). The finite-sample lower-confidence estimator of $\mathtt{D}_Z = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}})$ is

$$\widehat{\mathtt{D}}_{\mathrm{LCE}} \equiv \left[\widehat{\mathcal{J}}\left(\widehat{f}_{\mathrm{tr}}; S_P^{\mathrm{val}}, S_Q^{\mathrm{val}}\right) - \Gamma_{\hat{\delta}}\right]_+.$$

Definition 8 specifies a finite-sample lower-confidence estimator.

Theorem 3 (DV-Based Correction). *Let $Z = \mathcal{M}(X) + B$ with deterministic $\mathcal{M}$ and $B \sim \mathcal{N}(0, \Sigma_B)$, and let $\widetilde{Q}_{\mathcal{M}}$ be defined by (4). Assume $P_{\mathcal{M},B} \ll \widetilde{Q}_{\mathcal{M}}$. For any measurable $f : \mathbb{R}^d \to \mathbb{R}$ with $\mathbb{E}_{\widetilde{Q}_{\mathcal{M}}}[e^{f(Z)}] < \infty$, define*

$$\widehat{\mathtt{D}}_Z(f) \equiv \mathcal{J}\big(f; P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}}\big) = \mathbb{E}_{P_Z}[f(Z)] - \log \mathbb{E}_{\widetilde{Q}_{\mathcal{M}}}\big[e^{f(Z)}\big].$$

Let $\widehat{\mathtt{D}}_{\mathrm{LCE}}$ be the finite-sample lower-confidence estimator from Definition 8. Then:

(i) $0 \le \sup_f \widehat{\mathtt{D}}_Z(f) = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}})$.

(ii) *For every f, $\widehat{\mathtt{D}}_Z(f) \le \mathtt{D}_{\mathrm{KL}}\big(P_{\mathcal{M},B} \big\| \widetilde{Q}_{\mathcal{M}}\big) \equiv \mathtt{D}_Z$.*

(iii) *With probability at least $1 - \hat{\delta}$ (over the independent validation splits in Definition 8), $0 \le \widehat{\mathtt{D}}_{\mathrm{LCE}} \le \mathtt{D}_Z$.*

Theorem 4 (SWD-Based Correction). *Let $Z = \mathcal{M}(X) + B$ with deterministic $\mathcal{M}$ and $B \sim \mathcal{N}(0, \Sigma_B)$, and let $\widetilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$ be defined by (4), and let $\lambda_{\max}(\Sigma_Z)$ be the largest eigenvalue of Σ_Z. Assume $P_{\mathcal{M},B} \ll \widetilde{Q}_{\mathcal{M}}$ and $\Sigma_Z \succ 0$. Define*

$$\widehat{\mathtt{D}}_Z \equiv \frac{1}{2\lambda_{\max}(\Sigma_Z)} \mathrm{SW}_2^2\big(P_{\mathcal{M},B}, \widetilde{Q}_{\mathcal{M}}\big).$$

Then $0 \le \widehat{\mathtt{D}}_Z \le \mathtt{D}_Z$.

Theorems 3 and 4 lead to Corollary 1.

Corollary 1. *Let $\mathcal{M} : \mathcal{X} \mapsto \mathbb{R}^d$ be an arbitrary deterministic mechanism and $B \sim \mathcal{N}(0, \Sigma_B)$ such that $\mathtt{LogDet}(\mathcal{M}(X), B) = \beta$. Under the assumptions of Theorems 3 and 4, the perturbed mechanism $Z = \mathcal{M}(X) + B$ is PAC private with*

$$\mathtt{MI}(X;Z) \le \beta - \widehat{\mathtt{D}}_Z < \beta, \tag{8}$$

where $\widehat{\mathtt{D}}_Z > 0$ is obtained by Theorem 3 ($\widehat{\mathtt{D}}_Z(f)$) or Theorem 4. In addition, if $\widehat{\mathtt{D}}_Z = \widehat{\mathtt{D}}_{\mathrm{LCE}}$, then (8) holds with probability at least $1 - \hat{\delta}$.

Corollary 1 shows that accounting for non-Gaussianity through the correction term $\widehat{\mathtt{D}}_Z > 0$ yields $\mathtt{MI}(X;Z) \le \beta - \widehat{\mathtt{D}}_Z < \beta$, where the correction is obtained via DV-based correction or sliced Wasserstein correction. In practice, $\widehat{\mathtt{D}}_Z$ estimates the Gaussianity gap $\mathtt{Gap}_d$, capturing the saved privacy budget, which is particularly valuable for budget savings in mechanism composition. However, this budget-saving approach is post-hoc after Auto-PAC privatization. Appendix F in [60] provides additional discussions and interpretations. Next, we propose a new privacy framework enabling automatic optimal privacy budget implementation.

4 Residual-PAC Privacy

Recall that PAC Advantage Privacy (Definition 3) quantifies the amount of *privacy leaked by* $\mathcal{M}(X)$ in terms of the posterior advantage Δ_f^δ encountered by the adversary. Complementing this perspective, we introduce the notion of *posterior disadvantage* encountered by the adversary, which captures the amount of *residual privacy protection* that persists after leakage by $\mathcal{M}(X)$.

To formalize this residual protection, we first define the *intrinsic privacy* of a data distribution $\mathcal{D}$ relative to a fixed reference distribution $\mathcal{R}$ on $\mathcal{X}$ such that *(i)* $\mathrm{supp}(\mathcal{D}) \subseteq \mathrm{supp}(\mathcal{R})$ and *(ii)* the f-divergence $\mathtt{D}_f(\mathcal{D}\|\mathcal{R})$ is finite (when $\mathtt{D}_f$ is the KL-divergence, this means the entropy of $\mathcal{R}$ is finite; see Section 4.1 for the formal definition of Shannon/differential entropy). The intrinsic privacy is then defined based on f-divergence as

$$\mathtt{IntP}_f(\mathcal{D}) = -\mathtt{D}_f(\mathcal{D} \,\|\, \mathcal{R}),$$

where $\mathtt{D}_f(\mathcal{D}\|\mathcal{R})$ is the f-divergence between $\mathcal{D}$ and $\mathcal{R}$, quantifying how much $\mathcal{D}$ deviates from the reference $\mathcal{R}$. Intuitively, $-\mathtt{D}_f(\mathcal{D}\|\mathcal{R})$ rewards distributions that remain close to the "random guess" using $\mathcal{R}$, and by construction $\mathtt{IntP}_f(\mathcal{D}) \le 0$, attaining zero exactly when $\mathcal{D} = \mathcal{R}$.

Examples of $\mathcal{R}$. When $\mathcal{X}$ is bounded, $\mathcal{R}$ can be the uniform law $\mathcal{U}$ on $\mathcal{X}$. However, on an unbounded $\mathcal{X}$, the uniform reference $\mathcal{R} = \mathcal{U}$ has infinite volume $\int_{\mathcal{X}} dx = \infty$, potentially making $\mathtt{IntP}_f(\mathcal{D})$ vacuous or undefined. To avoid this, we instead require $\mathcal{R}$ to satisfy $\mathtt{D}_f(\mathcal{D}\|\mathcal{R}) < \infty$. For example, one can choose $\mathcal{R}$ by: *(i)* truncated uniform on a large but bounded region containing $\mathrm{supp}(\mathcal{D})$, *(ii)* maximum-entropy Gaussian matching known moments of $\mathcal{D}$, or *(iii)* smooth pullback of uniform on $(0,1)^d$ via bijection (e.g., componentwise sigmoid). Under any of these constructions, $\mathcal{R}$ retains the "random-guess" semantics yet has finite $\mathtt{D}_f(\mathcal{D} \,\|\, \mathcal{R})$, ensuring $\mathtt{IntP}_f(\mathcal{D})$ remains meaningful even on unbounded $\mathcal{X}$. Please see Appendix E in [60] for a detailed discussion.

Definition 9 ($(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC (R-PAC) Privacy). *A mechanism $\mathcal{M}$ is said to be* $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC (R-PAC) private *if it is $(\delta, \rho, \mathcal{D})$ PAC private and*

$$\mathtt{R}_f^\delta \equiv \mathtt{IntP}_f(\mathcal{D}) - \mathtt{D}_f(\mathbf{1}_\delta \| \mathbf{1}_{\delta_o^\rho}),$$

is the posterior disadvantage*, where $\mathbf{1}_\delta$ and $\mathbf{1}_{\delta_o^\rho}$ are indicator distributions representing the adversary's inference success before and after observing the mechanism's output, respectively.*

The posterior disadvantage $\mathtt{R}_f^\delta$ captures the *residual privacy guarantee*, which is the portion of intrinsic privacy (w.r.t. a reference $\mathcal{R}$) that remains uncompromised after the privacy loss $\Delta_f^\delta = \mathtt{D}_f(\mathbf{1}_\delta \| \mathbf{1}_{\delta_o^\rho})$ (Definition 3). Then, the total intrinsic privacy is precisely decomposed as

$$\mathtt{IntP}_f(\mathcal{D}) = \mathtt{R}_f^\delta + \Delta_f^\delta. \tag{9}$$

This relationship provides a complete and interpretable quantification of privacy risk, distinguishing between the privacy that is lost and that which endures after information disclosure via $\mathcal{M}(X)$. Analogous to PAC Privacy, membership inference

attacks (MIA) and R-PAC Membership Privacy can be instantiated from R-PAC Privacy. See Appendix C in [60] for detailed constructions.

4.1 Foundation of Residual-PAC Privacy

In this section, we develop general results to support concrete analyses under R-PAC Privacy framework. We begin by introducing key information-theoretic quantities, entropy and conditional entropy.

Entropy. The *Shannon entropy* of a discrete random variable X on alphabet $\mathcal{X}$ is given by

$$\mathcal{H}(X) = -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x)$$

while for continuous X, the *differential entropy* is

$$h(X) = -\int_{\mathcal{X}} f_X(x)\log f_X(x)\,dx.$$

Conditional Entropy. Let (X,W) be jointly distributed random variables. When X is discrete, the *conditional entropy* of X given W is defined by

$$\mathcal{H}(X|W) \equiv \mathbb{E}_W[\mathcal{H}(X|W=w)].$$

When X is continuous, the conditional entropy is $h(X|W) \equiv \mathbb{E}_W[h(X|W=w)]$. Here, the expectation is $\sum_{w\in\mathcal{W}} P_W(w)(\cdot)$ if W is discrete with mass P_W, and $\int_{\mathcal{W}} f_W(w)(\cdot)\,dw$ if W is continuous with density f_W.

For ease of exposition, we use $\mathcal{H}(X)$ to denote the entropy of X, either Shannon or differential depending on the context, and $\mathcal{H}(X|W)$ to denote the corresponding conditional entropy. When all entropies are finite, mutual information can equivalently be expressed as

$$\mathtt{MI}(X;W) = \mathcal{H}(X) - \mathcal{H}(X|W). \tag{10}$$

Consider any f-divergence $\mathtt{D}_f$, Theorem 1 of [56] shows that the posterior advantage Δ_f^δ is bounded by the minimum f-divergence between the joint distribution of $(X,\mathcal{M}(X))$, denoted by $P_{X,\mathcal{M}(X)}$, and the product of the marginal distribution P_X and any auxiliary output distribution P_W independent of X:

$$\Delta_f^\delta \;\le\; \inf_{P_W} \mathtt{D}_f\big(P_{X,\mathcal{M}(X)} \,\|\, P_X \otimes P_W\big), \tag{11}$$

where $P_{X,\mathcal{M}(X)}$ denotes the joint distribution of the data and mechanism output, $P_X = \mathcal{D}$, and P_W ranges over all distributions on the output space. When $\mathtt{D}_f$ is instantiated as $\mathtt{D}_{\mathrm{KL}}$ and $P_W = P_{\mathcal{M}(X)}$, we obtain (1).

Thus, for any f-divergence $\mathtt{D}_f$, inequality (11) implies that a mechanism $\mathcal{M}:\mathcal{X}\to\mathcal{Y}$ satisfies $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ R-PAC Privacy if

$$\mathtt{R}_f^\delta \;\ge\; \mathtt{IntP}_f(\mathcal{D}) - \inf_{P_W} \mathtt{D}_f\left(P_{X,\mathcal{M}(X)} \,\|\, P_X \otimes P_W\right). \tag{12}$$

Let R be a random variable of the reference $\mathcal{R}$ over $\mathcal{X}$. Corollary 2 follows from Theorem 1 of [56].

Corollary 2. *Suppose that $\mathcal{H}(X)$ is finite and let $\mathtt{D}_f$ be the KL divergence. A mechanism $\mathcal{M}:\mathcal{X}\to\mathcal{Y}$ satisfies $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ R-PAC Privacy if*

$$\mathtt{R}_f^\delta \ge \mathcal{H}(X|\mathcal{M}(X)) - \mathtt{V},$$

where $\mathtt{V} = \mathcal{H}(R)$ is the entropy of the reference distribution.

Corollary 2 establishes that when $\mathcal{H}(X)$ is finite, residual privacy $\mathtt{R}_f^\delta$ is lower bounded by $\mathcal{H}(X|\mathcal{M}(X)) - \mathtt{V}$, where $\mathtt{V}$ is independent of both data distribution $\mathcal{D}$ and mechanism $\mathcal{M}$. Since $\mathtt{V}$ is constant, $\mathtt{R}_f^\delta - \mathtt{V}$ effectively provides a privacy quantification lower-bounded by conditional entropy $\mathcal{H}(X|\mathcal{M}(X))$. If $\mathtt{D}_f(\mathcal{D}\|\mathcal{R}) < \infty$, then the inequality (12) holds without requiring $\mathcal{H}(X) < \infty$.

4.2 Stackelberg Residual-PAC Automatic Privatization

In this section, we present our algorithms for automatic R-PAC privatization when the f-divergence is instantiated with KL divergence, under which the worst-case residual privacy is quantified by conditional entropy. For a utility loss function $\mathcal{K}$, we define the optimal perturbation problem for any R-PAC privacy budget $\hat{\beta}$ as:

$$\inf_{Q} \mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B;\mathcal{M})] \quad \text{s.t.} \quad \mathcal{H}(X|\mathcal{M}(X)+B) \ge \hat{\beta},\ B\sim Q. \tag{13}$$

When $\mathcal{H}(X)$ is finite, by (10), any solution Q^* to problem (13) also solves (6) with PAC privacy budget $\beta = \mathcal{H}(X) - \hat{\beta}$. In addition, since $\mathtt{MI}(X;\mathcal{M}(X)+B) = \mathcal{H}(X) - \mathcal{H}(X|\mathcal{M}(X)+B)$ with finite $\mathcal{H}(X)$, solving the optimal perturbation problem (13) with conditional entropy constraints presents the same computational challenges as (6).

To address this limitation, we present a novel automatic privatization algorithm for R-PAC privacy, termed *Stackelberg Residual-PAC (SR-PAC)*. Our approach is based on a Stackelberg game-theoretic characterization of the optimization (13). We show that SR-PAC achieves optimal perturbation without wasting privacy budget. Consequently, when $\mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B;\mathcal{M})] = \mathbb{E}_Q[\|B\|_2^2]$, SR-PAC can achieve superior utility performance compared to Auto-PAC and Efficient-PAC (Appendix D in [60]) for the same mutual information privacy budget.

Our SR-PAC algorithm recasts the optimal perturbation problem (13) as a Stackelberg game between a *Leader* (who chooses the *perturbation rule* Q) and a *Follower* (who chooses the *decoder* attempting to infer X from Y). Let Γ denote a rich family of noise distributions. Let $\Pi = \{\pi : \pi(\cdot|y) \in \Delta(\mathcal{X}), y\in\mathcal{Y}\}$ denote a rich family of decoder distributions (e.g., all conditional density functions on $\mathcal{X}$ given $\mathcal{Y}$, or a parameterized neural network family).

Follower's Problem. For a fixed perturbation rule Q, the Follower chooses decoder π to minimize the expected log

Algorithm 2 Monte Carlo SR-PAC

Require: Privacy budget $\hat{\beta}$, decoder family Π_ϕ, perturbation rule family Γ_λ, utility loss $\mathcal{K}(\cdot)$, learning rates η_ϕ, η_λ, penalty weight σ, iterations T_λ, T_ϕ, batch size m

Initialize parameters $\lambda, \phi \sim \text{init}()$
for $t = 1, \ldots, T_\lambda$ **do**
  if $t \bmod T_\phi = 0$ **then**
    Update Decoder:
    for $i = 1, \ldots, T_\phi$ **do**
      Sample $\{(x_j, b_j, y_j)\}_{j=1}^m$ where $x_j \sim \mathcal{D}$, $b_j \sim Q_\lambda$, $y_j = \mathcal{M}(x_j) + b_j$
      $\widehat{W} = \frac{1}{m}\sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$
      $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \widehat{W}$
    end for
  end if
  Update Perturbation Rule:
  Sample $\{(x_j, b_j, y_j)\}_{j=1}^m$ where $x_j \sim \mathcal{D}$, $b_j \sim Q_\lambda$, $y_j = \mathcal{M}(x_j) + b_j$
  $H_c = \frac{1}{m}\sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$
  $\mathcal{L}_\lambda = \frac{1}{m}\sum_{j=1}^m \mathcal{K}(b_j) + \sigma(H_c - \hat{\beta})^2$
  $\lambda \leftarrow \lambda - \eta_\lambda \nabla_\lambda \mathcal{L}_\lambda$
end for
return Optimal parameters (λ^*, ϕ^*)

score

$$W(Q,\pi) \equiv \mathbb{E}_{X\sim\mathcal{D},B\sim Q}\left[-\log \pi(X|\mathcal{M}(X)+B)\right].$$

That is, the follower aims to find $\pi^*(Q) \in \arg\inf_{\pi\in\Pi} W(Q,\pi)$.

Leader's Problem. Given a privacy budget $\hat{\beta}$, the Leader chooses Q to solve

$$\inf_{Q\in\Gamma} \mathbb{E}_{X\sim\mathcal{D},B\sim Q}\left[\mathcal{K}(B;\mathcal{M})\right], \text{ s.t. } \inf_{\pi\in\Pi} W(Q,\pi) \geq \hat{\beta}.$$

Therefore, a profile (Q^*, π^*) is a *Stackelberg equilibrium* if it satisfies

$$\begin{cases} Q^* \in \arg\inf_{Q\in\Gamma} \mathbb{E}[\mathcal{K}(B;\mathcal{M})], \text{ s.t. } W\left(Q, \pi^*(Q)\right) \geq \hat{\beta}, \\ \pi^*(Q) \in \arg\inf_{\pi\in\Pi} W(Q,\pi). \end{cases} \tag{14}$$

When we consider output perturbation and the utility loss $\mathcal{K}$ is chosen such that $Q \mapsto \mathbb{E}_{X\sim P_X, B\sim Q}\left[\mathcal{K}(B;\mathcal{M})\right]$ is convex in Q, the problem (14) is convex in both Q and π. Specifically, for each fixed perturbation rule Q, the map $\pi \mapsto W(Q,\pi)$ is a convex function of π. Similarly, for each fixed decoder π, the function $Q \mapsto W(Q,\pi)$ is convex in Q. Because these two convexity properties hold simultaneously, $(Q,\pi) \mapsto W(Q,\pi)$ is jointly convex on $\Gamma \times \Pi$. By the partial minimization theorem [6, Section 3.2.5], taking the pointwise infimum over π preserves convexity in Q. Thus, $Q \mapsto \inf_{\pi\in\Pi} W(Q,\pi)$ is a convex function of Q. Consequently, once the Follower replaces π by its best response $\pi^*(Q)$, the Leader's feasible set $\{Q \in \Gamma : \inf_{\pi\in\Pi} W(Q,\pi) \geq \hat{\beta}\}$ is convex, and minimizing the convex utility loss function $Q \mapsto \mathbb{E}_{X\sim P_X, B\sim Q}\left[\mathcal{K}(B;\mathcal{M})\right]$ over this set remains a convex program in Q. Meanwhile, the Follower's problem $\inf_{\pi\in\Pi} W(Q,\pi)$ is convex in π for any fixed Q. Thus, the Stackelberg game reduces to a single-level convex optimization in Q, with the inner decoder problem convex in π.

Proposition 5 shows that the Stackelberg equilibrium perturbation rule solves (13).

Proposition 5. *Let (Q^*, π^*) be a Stackelberg equilibrium satisfying (14) for any given $\hat{\beta}$. Then, Q^* solves (13) with privacy budget $\hat{\beta}$. In addition, in any Stackelberg equilibrium (Q^*, π^*), $\pi^* = \pi^*(Q^*)$ is unique.*

Algorithm 2 provides a Monte-Carlo-based approach to solve the Stackelberg equilibrium (14). By Monte Carlo sampling, the algorithm trains the decoder by minimizing reconstruction loss on perturbed data, allowing it to adapt to the current noise distribution. It then updates the perturbation rule by minimizing utility loss subject to the privacy constraint, implemented via a penalty term that drives the privacy cost toward the target budget. For scalability, the online extended version [60] (Appendix H) also presents two variants, *Sliced R-PAC Privacy* and *Sliced SR-PAC* algorithm, based on *sliced mutual information* [24]. The online extended version [60] also provides finite-sample and approximate-optimization error analyses for the Follower (Appendix G.1).

5 Properties of SR-PAC Privatization

This section presents some important properties of SR-PAC.

5.1 Anisotropic Noise Perturbation

The Auto-PAC perturbs the mechanism using *anisotropic* Gaussian noise as much as needed in each direction of the output. This direction-dependent noise addition yields better privacy-utility tradeoffs than isotropic perturbation. SR-PAC also supports anisotropic perturbation under Assumption 1.

Assumption 1. *For an arbitrary deterministic mechanism $\mathcal{M}$, we assume the following.*

(i) Every $Q \in \Gamma$ is log-concave.

(ii) For any orthonormal direction $w \in \mathbb{R}^d$, $\langle \mathcal{M}(X), w\rangle$ is non-degenerate.

(iii) The utility function $\mathcal{K}$ is radial (depends only on $\|B\|_2$) and strictly convex in the eigenvalues of covariance matrix Σ_Q of Q. For example, $\kappa(B) = \|B\|_2^2$.

(iv) There exist orthonormal $u, v \in \mathbb{R}^d$ such that the marginal entropy gain per unit variance along u exceeds that along v. That is, for any $\sigma^2 >$

0, $\frac{\partial}{\partial \sigma_u^2}\mathcal{H}(X|Z_u)|_{\sigma^2} > \frac{\partial}{\partial \sigma_v^2}\mathcal{H}(X|Z_v)|_{\sigma^2}$, *where* $Z_w = \mathcal{M}_w(X) + B_w$, *with* $A_w(X) = \langle A(X), w\rangle$ *for* $A \in \{\mathcal{M}, B\}$, $w \in \{u, v\}$.

Assumption 1 ensures that SR-PAC's optimization is convex and admits a genuinely anisotropic solution: requiring each noise distribution in Γ to be log-concave makes the feasible set convex and tractable; non-degeneracy of $< \mathcal{M}(X), w >$ for every unit vector w guarantees that every direction affects information leakage; a strictly convex, radial utility K yields a unique cost-to-noise mapping; and the existence of two orthonormal directions whose marginal entropy gain per unit variance differs implies that allocating noise unevenly strictly outperforms isotropic noise.

Proposition 6. *Under Assumption 1, any Stackelberg equilibrium perturbation rule Q^* is anisotropic. That is, its covariance matrix Σ_{Q^*} satisfies*

$$r_{\max}(\Sigma_{Q^*}) > r_{\min}(\Sigma_{Q^*}),$$

where $r_{\max}(\Sigma_{Q^*})$ and $r_{\min}(\Sigma_{Q^*})$ are the maximum and the minimum eigenvalues of Σ_{Q^*}.

Proposition 6 demonstrates that SR-PAC allocates noise exclusively to privacy-sensitive directions, with high-leakage dimensions receiving proportionally more noise than low-leakage dimensions. This targeted approach achieves desired privacy levels with minimal total perturbation, preserving task-relevant information with reduced noise.

5.2 Directional-Selectivity of SR-PAC

Let Z be a d-dimensional real-valued *output vector* produced by a deterministic mechanism $\mathcal{M}(X)$. Throughout we assume $\Sigma_Z \succ 0$ and finite differential entropy $\mathcal{H}(Z)$. For any application, let $S_{\text{task}} \subseteq \mathbb{R}^d$ denote a practitioner-chosen *task-critical sub-space* (the directions whose preservation matters most) and write Π_{task} for the orthogonal projector onto it.

Classification tasks. In what follows we illustrate the theory with multi-class classification, where Z is the *logit* vector, $\hat{y} = \arg\max_i Z_i$, and $S_{\texttt{lab}} \equiv \text{span}\{e_\ell - e_j : j \neq \ell\}$, where $\texttt{lab}$ means "label". Let $\Pi_{\texttt{lab}}$ be the projector onto $S_{\texttt{lab}}$. The analysis for a general S_{task} is identical after replacing $\texttt{lab}$ by task.

For any privacy budget $0 < \beta < \mathcal{H}(Z)$, consider Q^* that solves

$$\inf_{Q:\texttt{MI}(Z;Z+B)=\beta} \mathbb{E}[\|B\|_2^2].$$

For every unit vector w, let $g(w) \equiv \frac{1}{2}\texttt{mmse}(\langle Z, w\rangle)$, where $\texttt{mmse}(\langle Z, w\rangle) \equiv \mathbb{E}\Big[\langle Z, w\rangle - \mathbb{E}\big[\langle Z, w\rangle | Y\big]\Big]^2$ is the *minimum mean-squared error* of estimating the scalar random variable $\langle Z, w\rangle$ from the noisy observation $Y = Z + B$.

Proposition 7. *Suppose $\mathcal{H}(Z)$ is finite. Fix any $0 < \beta < \mathcal{H}(Z)$. The following holds.*

(i) Let $\mathcal{N}(0, \Sigma_{\text{PAC}})$ be the Gaussian noise distribution used by the Auto-PAC such that $\texttt{LogDet}(Z, B_{\text{PAC}}) = \beta$. If Z is non-Gaussian, then $\mathbb{E}_{Q^*}[\|B\|_2^2] < \mathbb{E}[\|B_{\text{PAC}}\|_2^2]$.

(ii) Suppose $\sup_{v \in S_{\texttt{lab}}, \|v\|=1} g(v) < \inf_{w \perp S_{\texttt{lab}}, \|w\|=1} g(w)$. Let $\beta_{\texttt{lab}} \equiv \frac{1}{2}\int_{w \perp S_{\texttt{lab}}} g(w) d\sigma_w^2$ denote the largest privacy budget that can be satisfied using noise supported entirely on $S_{\texttt{lab}}^{\perp}$. Then, for every $\beta \leq \beta_{\texttt{lab}}$, we have $\Pi_{\texttt{lab}} B^* = 0$ a.s., $\arg\max_i (Z_i + B_i^*) = \hat{y}$ a.s.

In Proposition 7, part (i) shows that SR-PAC always uses strictly less noise magnitude than any Auto-PAC (regardless of how anisotropic the Auto-PAC noise covariance may be) because Auto-PAC treats Z as Gaussian and thus overestimates the required variance when Z is non-Gaussian. Part (ii) demonstrates that, under the natural ordering of directional sensitivities, SR-PAC allocates its noise budget exclusively in directions orthogonal to the label sub-space until a critical threshold $\beta_{\texttt{lab}}$ is reached. In practice, this means SR-PAC perturbs only "utility-harmless" dimensions first, preserving the predicted class and concentrating protection where it is most needed, thereby outperforming Auto-PAC in any scenario where certain directions leak more information than others.

5.3 Sensitivity to β

Sensitivity to the privacy parameter β is crucial for predictable and accurate control of privacy-utility trade-off. Let $\texttt{Priv}_\beta$ and $\texttt{Util}_\beta$, respectively, denote the sensitivities of privacy and utility (for certain measures). If $\texttt{Priv}_\beta = 1$, then any infinitesimal increase $\Delta\beta$ in the privacy budget raises the true mutual information $\texttt{MI}(X;Y)$ by exactly $\Delta\beta$. Thus, no part of the privacy budget is "wasted" or "over-consumed". By contrast, if $\texttt{Priv}_\beta < 1$, then increasing β may force additional noise without achieving the full allowed leakage; and if $\texttt{Priv}_\beta > 1$, then increasing the budget by $\Delta\beta$ can increase the true leakage by more than $\Delta\beta$. In particular, if the mechanism is calibrated to be tight at β (i.e., $\texttt{MI}(X;Y) = \beta$), then it may become over-budget, i.e., $\texttt{MI}(X;Y) > \beta + \Delta\beta$. Similarly, if $\texttt{Util}\beta$ is high, then an infinitesimal increase $\Delta\beta$ in the privacy budget yields a large improvement in utility; if $\texttt{Util}_\beta$ is low, the same increase yields a small improvement, indicating inefficient conversion of the privacy budget into utility gains.

Let $V_{\text{SR}}(\beta) \equiv \min_{Q:\texttt{MI}(X;\mathcal{M}(X)+B)\leq\beta} \mathbb{E}_Q\big[\|B\|_2^2\big]$ be the optimal noise-power curve attained by SR-PAC, and let $\texttt{MI}_{\text{SR}}(\beta)$ as the corresponding true mutual information attained by SR-PAC. Let $V_{\text{PAC}}(\beta) \equiv \text{tr}(\Sigma_{B_{\text{PAC}}}(\beta))$, where $Q(\beta) = \mathcal{N}(0, \Sigma_{B_{\text{PAC}}}(\beta))$ solves $\texttt{LogDet}(\mathcal{M}(X), B_{\text{PAC}}) = \beta$. In addition, let $\texttt{MI}_{\text{PAC}}(\beta) \equiv \beta - \texttt{Gap}_{\texttt{d}}(Q(\beta))$, where $\texttt{Gap}_{\texttt{d}}(Q) = \texttt{D}_{\text{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}})$ with $B \sim Q$, and $\widetilde{Q}_{\mathcal{M}}$ given by (4). Define $\texttt{Priv}_\beta^{\text{SR}} \equiv \frac{d}{d\beta}\texttt{MI}_{\text{SR}}(\beta)$, $\texttt{Priv}_\beta^{\text{PAC}} \equiv \frac{d}{d\beta}\texttt{MI}_{\text{PAC}}(\beta)$, $\texttt{Util}_\beta^{\text{SR}} \equiv \frac{d}{d\beta}(-V_{\text{SR}}(\beta))$, and $\texttt{Util}_\beta^{\text{PAC}} \equiv \frac{d}{d\beta}(-V_{\text{PAC}}(\beta))$.

Theorem 5. *For any data distribution $\mathcal{D}$, let $\mathcal{M}$ be an arbitrary deterministic mechanisms such that $\mathcal{M}(X)$ is non-Gaussian with $\Sigma_M \succ 0$. The following holds.*

(i) $\texttt{Priv}_\beta^{\text{PAC}} \leq \texttt{Priv}_\beta^{\text{SR}} = 1$, *with strict inequality for non-Gaussian $\mathcal{M}(X)$.*

(ii) $\texttt{Util}_\beta^{\text{SR}} \geq \texttt{Util}_\beta^{\text{PAC}}$, *with equality only for Gaussian $\mathcal{M}(X)$.*

Theorem 5 proves that SR-PAC with arbitrary noise distributions achieves: (i) *Exact leakage-budget alignment* ($\texttt{Priv}_\beta^{\text{SR}} = 1$), (ii) *Stricter utility decay* for Auto-PAC ($\texttt{Util}_\beta^{\text{SR}} \geq \texttt{Util}_\beta^{\text{PAC}}$). This holds for all non-Gaussian $\mathcal{M}(X)$ when increasing privacy strength (i.e., β decreasing). A robustness analysis under finite-sample calibration and optimization effects is given in Appendix G.2 of [60].

5.4 Composition

Graceful composition properties in privacy definitions like DP make privacy loss quantifiable under multiple operations on datasets. This enables modular system design: each component can be tuned to a local privacy–utility tradeoff, while composition rules provide an explicit bound on the overall (global) privacy risk. Consider k mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_k$, where each $\mathcal{M}_i(\cdot, \theta_i) : \mathcal{X} \mapsto \mathcal{Y}_i$ with $\theta_i \in \Theta_i$ as the random seed. Let $\vec{\mathcal{Y}} = \prod_{i=1}^k \mathcal{Y}_i$ and let $\vec{\Theta} = \prod_{i=1}^k \Theta_i$. The composition $\overrightarrow{\mathcal{M}}(\cdot, \vec{\theta}) : \mathcal{X} \mapsto \vec{\mathcal{Y}}$ is defined as $\overrightarrow{\mathcal{M}}(X, \vec{\theta}) = (\mathcal{M}_1(X, \theta_1), \ldots, \mathcal{M}_k(X, \theta_k))$. PAC Privacy composes gracefully [56]. In particular, for independent mechanisms applied to the same dataset, mutual information bounds compose additively: if each $\mathcal{M}_i$ is PAC Private with bound β_i, then $\overrightarrow{\mathcal{M}}$ has bound $\sum_{i=1}^k \beta_i$.

R-PAC Privacy also enjoys additive composition with respect to conditional entropy bounds. Suppose each mechanism $\mathcal{M}_i$ is R-PAC private with conditional entropy lower bound $\hat{\beta}_i$. By (10), this implies that $\mathcal{M}_i$ is PAC private with privacy budget $\beta_i = \mathcal{H}(X) - \hat{\beta}_i$. Then, by Theorem 7 of [56], the composition $\overrightarrow{\mathcal{M}}(X, \vec{\theta})$ is PAC private with total mutual information upper bounded by $\sum_{i=1}^k (\mathcal{H}(X) - \hat{\beta}_i)$. Equivalently, the composition $\overrightarrow{\mathcal{M}}(X, \vec{\theta})$ is R-PAC private with overall conditional entropy lower bounded by $\sum_{i=1}^k \hat{\beta}_i - (k-1)\mathcal{H}(X)$.

However, this additive composition property for mutual information yields conservative aggregated privacy bounds [56], and utility degradation compounds when each mechanism $\mathcal{M}_i$ uses conservative privacy budgets β_i. To address this limitation, we employ an optimization-based approach within the SR-PAC framework to compute tighter conditional entropy bounds. Consider k mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k$, where each $\mathcal{M}_i$ is privatized by the perturbation rule Q_i to satisfy R-PAC privacy with bounds $\hat{\beta}_i$. The Leader designs these perturbation rules $Q_1, \ldots, Q_k$, while the Follower finds the optimal decoder for the joint composition $\overrightarrow{\mathcal{M}}(X, \vec{\theta}) = (\mathcal{M}_1(X), \ldots, \mathcal{M}_k(X))$: $\inf_{\pi \in \Pi} W(\pi; \overrightarrow{\mathcal{M}}) \equiv \mathbb{E}_{X \sim \mathcal{D}} \left[-\log \pi(X | \overrightarrow{\mathcal{M}}(X), \vec{\theta}) \right]$. This game-theoretic formulation allows for tighter privacy-utility tradeoffs in composed systems by optimizing the joint privatization strategy. This joint SR-PAC formulation also extends to adaptive composition, where each Q_i (and the corresponding decoder update) may be chosen sequentially based on previously released privatized outputs, in the same spirit as the adaptive composition procedure of PAC Privacy [56].

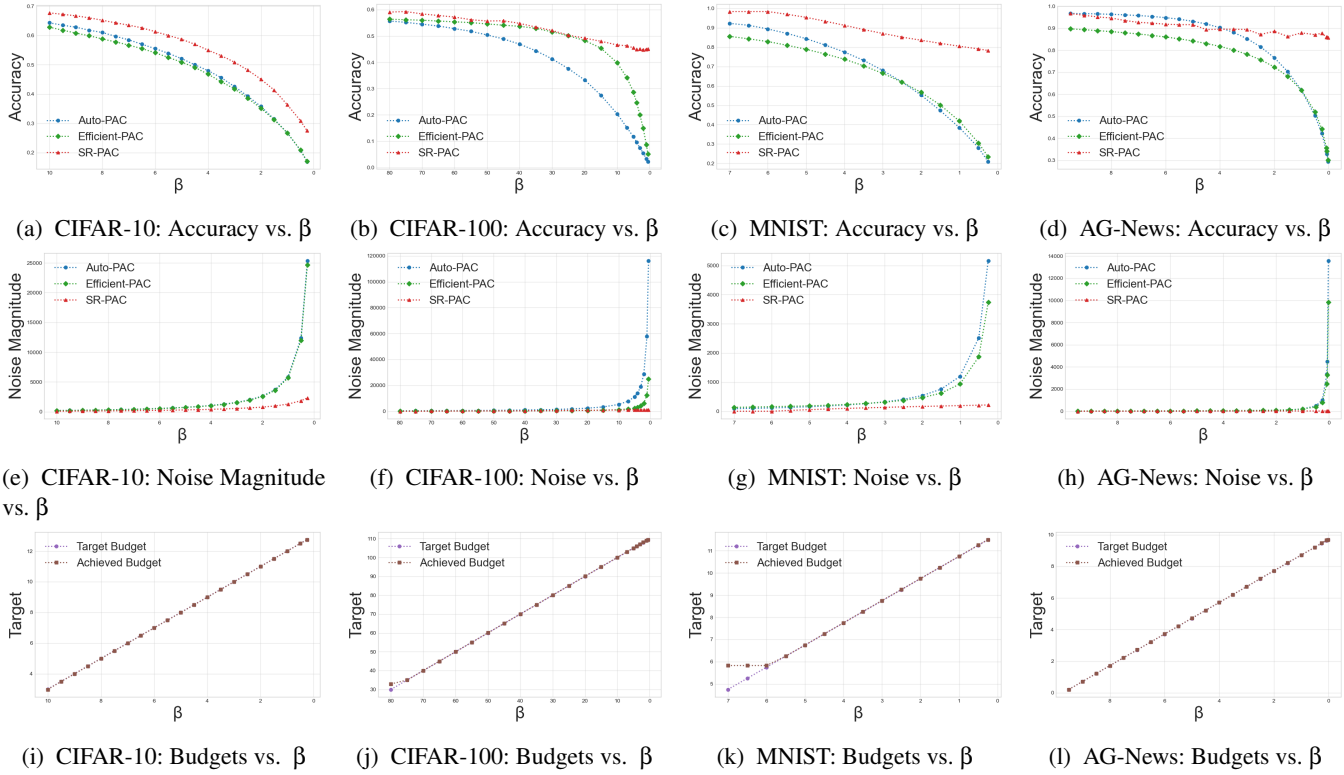
6 Experiments

We conduct two sets of experiments to evaluate our approach. First, we compare SR-PAC against Auto-PAC and Efficient-PAC (Appendix D in [60]) using CIFAR-10 [35], CIFAR-100 [35], MNIST [37], and AG-News [62] datasets, with results presented in Section 6.1. We use (R-)PAC to refer to the family of SR-PAC, Auto-PAC, and Efficient-PAC. Second, we extend this comparison to include DP by equalizing optimal posterior success rates of membership inference (Appendix A.1) across all methods, making their privacy budgets comparable. For this comparison, we use Iris [21] and Rice [12] datasets, with results shown in Section 6.2. All experiments focus on output perturbation. Appendix U in [60] provides more details.

CIFAR-10 and Base Classifier. CIFAR-10 and base classifier. We evaluate on CIFAR-10 (32×32 RGB, 10 classes) with standard per-channel normalization (mean 0.5, std 0.5). The unperturbed classifier is a small CNN with two Conv–ReLU–MaxPool blocks (2×2 pooling; 32 and 64 channels), followed by a 128-unit fully connected ReLU layer and a 10-logit output layer. It is trained with cross-entropy loss, and predicts the argmax logit at inference. The unperturbed classifier achieves 0.7181 ± 0.0088 accuracy.

CIFAR-100 and Base Classifier. We evaluate on CIFAR-100 (32×32 RGB, 100 classes) with standard per-channel normalization (mean 0.5, std 0.5). The unperturbed classifier is a deeper CNN with three convolutional blocks (two 3×3 Conv–BatchNorm–ReLU layers per block, followed by 2×2 max-pooling), with channel widths 64/128/256. A three-layer MLP head (4096→512→256→100) with ReLU and dropout (0.5) outputs 100 logits, and prediction is by argmax. The unperturbed classifier achieves 0.5914 ± 0.0090 accuracy.

MNIST dataset and Base Classifier. We evaluate on MNIST (60,000 train / 10,000 test) consisting of 28×28 grayscale digit images. Each image is loaded as a $1 \times 28 \times 28$ tensor and normalized per channel (mean 0.1307, std 0.3081). The unperturbed classifier is a CNN with two Conv2d→BatchNorm→ReLU→MaxPool (2×2) blocks (channels $1 \to 32 \to 64$), yielding a $64 \times 7 \times 7$ feature map, followed by a two-layer fully connected head (128 units with ReLU+Dropout, then 10 logits). At inference it outputs a 10-dimensional logit vector and predicts by argmax. The unperturbed classifier achieves 0.9837 accuracy.

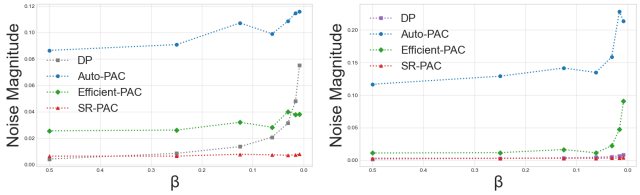


(a) CIFAR-10: Accuracy vs. β (b) CIFAR-100: Accuracy vs. β (c) MNIST: Accuracy vs. β (d) AG-News: Accuracy vs. β

(e) CIFAR-10: Noise Magnitude vs. β (f) CIFAR-100: Noise vs. β (g) MNIST: Noise vs. β (h) AG-News: Noise vs. β

(i) CIFAR-10: Budgets vs. β (j) CIFAR-100: Budgets vs. β (k) MNIST: Budgets vs. β (l) AG-News: Budgets vs. β

Figure 1: Empirical comparisons of SR-PAC, Auto-PAC (Algorithm 1), and Efficient-PAC (Algorithm 3 in [60]) on CIFAR-10, CIFAR-100, MNIST, and AG-News as β varies. Each column corresponds to one dataset; within each column, the three panels report (top) classification accuracy of the perturbed model versus the target budget β, (middle) the average noise magnitude $\mathbb{E}[\|B\|_2^2]$ used by each method, and (bottom) the "target versus achieved" privacy budget (conditional entropy) for our SR-PAC.



(a) Iris (b) Rice

Figure 2: Empirical comparisons of DP, Auto-PAC, Efficient-PAC, and SR-PAC on mean estimations, using Iris and Rice datasets, in terms of average noise magnitude $\mathbb{E}[\|B\|_2^2]$. All the numerical values are shown in Tables 5 and 6.

AG-News dataset and Base Classifier. We evaluate on AG-News with 120,000 training and 7,600 test articles evenly split across four classes (World, Sports, Business, Sci/Tech), i.e., 30,000 training and 1,900 test examples per class. Each example's title and description are concatenated, lowercased, and tokenized by whitespace (truncated/padded to 64 tokens). We build a 30,000-word vocabulary from the training split, map tokens to indices (out-of-vocabulary as 0), and feed the indices into an `nn.EmbeddingBag` layer (embedding size 300, mean-pooling) to obtain a 300-dimensional document vector. A two-layer MLP head (300 → 256 with ReLU and 0.3 dropout, then 256 → 4) produces a 4-dimensional logit vector, and prediction is by argmax. The unperturbed mechanism achieves 0.9705 accuracy.

Recall that β upper-bounds $\mathtt{MI}\big(X;\mathcal{M}(X)+B\big)$, and SR-PAC enforces the equivalent constraint $\mathcal{H}\big(X \mid \mathcal{M}(X)+B\big) \geq \widehat{\beta} = \mathcal{H}(X) - \beta$. Although $\mathcal{H}(X)$ is unknown, we estimate for the purpose of evaluation to verify the tightness of the privacy bounds. Let $\mathtt{MI}_0 = \mathtt{MI}\big(X;\mathcal{M}(X)\big)$. By data processing inequality, $\mathtt{MI}\big(X;\mathcal{M}(X)+B\big) \leq \mathtt{MI}_0$ for any independent B, so the feasible budgets are $0 < \beta \leq \mathtt{MI}_0$ and this interval is common to Auto-PAC, Efficient-PAC, and SR-PAC. At $\beta = \mathtt{MI}_0$, the optimal choice is $B = 0$, and all methods coincide at the noiseless accuracy. This shared endpoint and feasible domain ensure that comparisons at a common target β are well-defined even without the exact $\mathcal{H}(X)$. Moreover, reparameterizing by achieved mutual information preserves the endpoint and domain, and, together with the small budget errors observed in panels (i–l), does not affect our empirical ordering. Under additive ℓ_2 output noise, the ordering by total noise magnitude $\mathbb{E}\left[\|B\|_2^2\right]$ coincides with the ordering by accuracy, consistent with the ℓ_2-based behavior reported in prior

work. Hence, the accuracy– and noise–vs.–β panels convey the same conclusion in our experiments.

6.1 (R-)PAC Comparison

For each dataset and its pretrained base classifier $\mathcal{M}$, we plot (i) the test accuracy of the perturbed model as a function of β, (ii) the average noise magnitude $\mathbb{E}[\|B\|_2^2]$ required to achieve each β, and (iii) SR-PAC's ability to hit the target $\mathcal{H}(X) - \beta$.

Accuracy vs. β (a–d of Figure 1). As β decreases (moving right), privacy increases and all methods lose test accuracy. For large β (near the no-privacy case), all three algorithms attain accuracies close to the noiseless model. As β tightens, the SR-PAC curve remains strictly above the Auto-PAC and Efficient-PAC curves across datasets. On **CIFAR-10** and **CIFAR-100**, Auto-PAC and Efficient-PAC are visibly separated from each other (not merely from SR-PAC), reflecting their different Gaussian calibrations. On **MNIST** and **AG-News**, the three methods cluster near the top for larger β, but SR-PAC retains a measurable accuracy edge at matched β.

Noise magnitude vs. β (e–h of Figure 1). As β decreases, each algorithm must add more noise, so all three curves rise. Across all datasets, SR-PAC uses the smallest $\mathbb{E}\left[\|B\|_2^2\right]$ at each β. Auto-PAC and Efficient-PAC both overshoot—they inject more noise than SR-PAC at matched β—and on CIFAR-100, MNIST and AG-News, they diverge from each other as well.

The empirical ordering in both accuracy and noise magnitude matches Theorem 5. Moreover, Figure 1 (c–d, g–h) exhibits the behavior predicted by Proposition 7 on **MNIST** and **AG-News**: for $\beta \leq \beta_{\mathtt{lab}}$, SR-PAC allocates noise predominantly in directions (approximately) orthogonal to the label subspace, preserving the predicted class over a wide budget range. Concurrently, its total noise remains substantially smaller than Auto-PAC and Efficient-PAC, whose conservative Gaussian calibrations overestimate the required variance on heavy-tailed (non-Gaussian) logits.

Budgets vs. β (i–l of Figure 1). These panels plot SR-PAC's target privacy budgets in terms of mutual-information bounds β (horizontal) against the achieved empirical conditional-entropy budget (vertical). In every dataset, the red points lie tightly along the $y = x$ line, confirming that SR-PAC solves its follower problem with high accuracy and enforces the desired privacy level with negligible budget error. This provides a reliable, data-driven guarantee that the privacy constraint is satisfied.

6.2 Comparison with Differential Privacy

We calibrate DP and (R-)PAC to the same (optimal) posterior success rate for membership inference attacks, then compare their utility in terms of noise magnitudes (i.e., ℓ_2-norm of the difference between original and perturbed outputs). The base mechanism is a mean estimator. Appendix A.1 provides the conversions between (optimal) posterior success rates, DP parameters (DP-to-posterior mapping), and mutual information budgets (MI-to-posterior mapping). Concretely, for DP we select $(\varepsilon, \bar{\delta})$ that yields the target posterior bound via the DP-to-posterior mapping, and for (R-)PAC we choose the budget β that yields the same posterior via the MI-to-posterior mapping; with subsampling rate $r = 0.5$ we have prior $p = 0.5$. In each trial, we construct a membership vector $m \in \{0,1\}^P$ by i.i.d. Bernoulli(0.5) draws, so the member count $S = \sum_i m_i$ is random. We follow similar treatments for DP as Section 6.3 of [53]: the DP baseline clips each row in ℓ_2 to radius C, adds Gaussian noise to the clipped sum, and divides by S to produce the privatized mean; (R-)PAC injects noise calibrated to β. We report $\mathbb{E}\left[\|B\|_2^2\right]$ at matched posterior success rates. In our output-perturbed mean setting, this quantity equals the expected squared ℓ_2 error of the released statistic. Hence the ordering by $\mathbb{E}\left[\|B\|_2^2\right]$ is identical to the ordering by ℓ_2 accuracy. Appendix A.1 gives detailed DP vs. (R-)PAC discussion.

Figure 2. On the Iris and Rice mean–estimation tasks, SR-PAC attains the smallest average noise magnitude $\mathbb{E}\left[\|B\|_2^2\right]$ across privacy budgets β. As β decreases (stricter privacy), the noise required by Auto-PAC and Efficient-PAC rises much more steeply, whereas SR-PAC grows gently; see the zoomed view in Fig. 3 (Appendix U in [60]). The DP baseline remains well above SR-PAC and, at small budgets on Iris, also exceeds Efficient-PAC. Appendix U in [60] further reports empirical membership–inference results for SR-PAC, DP, Auto-PAC, and Efficient-PAC on these privatized mechanisms.

Auto-PAC and Efficient-PAC allocate *anisotropic* noise, but their shapes are task-agnostic and depend only on second-order structure, via covariance scaling (Auto-PAC) or eigen–allocation (Efficient-PAC). In small-sample regimes (e.g., Iris and Rice), the covariance spectrum is noisy and often ill-conditioned, and these moment-based rules propagate that instability into the noise design, yielding conservative noise levels, especially for small β. By contrast, SR-PAC enforces the conditional-entropy budget directly, leading to tighter budget implementation and lower required noise. Empirically (Figure 2), SR-PAC attains smaller average noise magnitudes across β with smoother scaling.

7 Conclusion

This work introduced R-PAC Privacy, an enhanced framework that guarantees privacy beyond Gaussian assumptions while overcoming the conservativeness of existing PAC Privacy algorithms. Our SR-PAC algorithm casts the privacy–utility trade-off as a Stackelberg problem, efficiently using the privacy budget and learning data- and mechanism-specific anisotropic noise. Extensive experiments show that SR-PAC consistently attains tighter privacy guarantees and higher utility than prior approaches, providing a rigorous and practical foundation for scalable privacy assurance in complex applications.

Ethical Considerations

We propose Residual-PAC Privacy (R-PAC) and its privatization scheme Stackelberg Residual-PAC (SR-PAC) as a privacy protection framework. Our goal is to reduce the conservativeness of prior PAC Privacy mechanisms so that, for a fixed privacy budget, practitioners can achieve better utility without relaxing stated privacy guarantees. All experiments use standard, publicly available benchmark datasets. We do not collect new data, interact with live production systems or APIs, scrape data, or discover/disclose vulnerabilities. We organize ethical considerations around the generic data-processing pipeline

$$\textbf{Data} \rightarrow \textbf{Mechanisms} \rightarrow \textbf{Downstream Decision-Making},$$

and assess benefits and harms using the Menlo Report principles [14]: Beneficence, Respect for Persons, Justice, and Respect for Law and Public Interest. Appendix B in [60] provides ethical considerations about the trade-offs across different privacy standards.

Stakeholders

Data are collections of records about people (e.g., medical, financial, behavioral). In our experiments, we use widely adopted public benchmarks, but in practice similar mechanisms could be deployed on sensitive real-world data. Relevant stakeholders include data subjects and dataset curators.
Mechanisms are data-driven systems (e.g., statistical analyses, ML models) operating under R-PAC or other frameworks (e.g., DP). Stakeholders include the privacy research community, data scientists/ML engineers, privacy/security teams, and auditors/defenders who pressure-test privacy claims.
Downstream decision-making comprises automated or human decisions relying on mechanism outputs (e.g., risk scoring, recommendation, decision support). Stakeholders include affected individuals (patients, applicants, platform users), as well as regulators, standards bodies, and advocacy organizations that rely on formal privacy statements.
Cross-cutting societal stakeholders (taxpayers, communities broadly subject to algorithmic systems, environmental stakeholders) are indirectly affected by how privacy-preserving data analysis becomes common practices.

Harms and Mitigations

Data stage. The primary privacy risk arises from how data are used, not from their mere existence. R-PAC does not introduce new collection/scraping/linkage activity; rather, it calibrates and interprets residual privacy risk for data use in mechanisms. A key harm is *misaligned expectations*: curators or deployers may choose parameters that are too weak for a given context, or may overgeneralize scenario-specific guarantees, creating a false sense of protection. We mitigate this by restricting experiments to public benchmarks and by emphasizing that guarantees are scenario-specific and depend on the chosen prior, model, and calibration. Real deployments require context-aware parameter selection and transparent communication of scope and limits.
Mechanism stage. The primary harm arises from the inevitable privacy–utility trade-off: stronger privacy typically requires injecting randomness that reduces output fidelity. Additional risks include (i) miscalibration or misuse from incorrect priors, assumptions, or implementations; (ii) privacy-washing or miscommunication if residual-risk metrics are presented as context-free guarantees; and (iii) increased engineering burden that raises the risk of implementation errors, especially for less-resourced teams. We provide finite-sample analysis and recommend documenting priors, modeling, and calibration choices, using conservative settings when assumptions are uncertain, and treating R-PAC bounds as one input to broader privacy/security review that includes independent auditing and pressure-testing.
Downstream stage. Noise can degrade decision quality and cause harmful errors in high-stakes contexts (e.g., healthcare, finance, admissions, risk scoring). Residual-risk metrics may also be overly relied upon as blanket approval for aggressive data use, even in domains where any non-zero residual risk or utility degradation is unacceptable. Our work does not deploy real systems; we use benchmarks to study trade-offs. For high-stakes deployments, we recommend conservative parameters, domain-specific evaluation of decision quality, and governance processes that do not treat formal bounds as the sole determinant of acceptability.
Cross-cutting societal impacts. If widely adopted, R-PAC-style methods may help align formal claims with empirical behavior and make residual risk explicit, supporting more realistic interpretation by auditors and regulators. However, they could be misused to justify aggressive data use if assumptions are obscured. Clear documentation, independent auditing, and appropriate oversight are important safeguards.

Decision to Conduct and Publish This Work

From a Beneficence perspective, our goal is to improve utility *at a fixed residual-privacy budget*, reducing harms associated with overly conservative or difficult-to-use mechanisms that can make formal privacy less practical. From a Respect for Persons and Justice perspective, making residual risk explicit and tying guarantees to concrete modeling and calibration choices supports more truthful, context-aware communication of privacy properties, while avoiding claims of absolute protection. From a Respect for Law and Public Interest perspective, more transparent and realistic privacy accounting can help regulators, standards bodies, and auditors evaluate systems where privacy claims depend on explicitly stated assumptions such as priors and model classes. We consider it ethically justified to conduct and publish this work, but deployment remains context-dependent.

Open Science

All artifacts necessary to evaluate our contribution consist solely of source code, available at the repository on Zenodo: https://doi.org/10.5281/zenodo.17871622.

The repository contains implementations of SR-PAC and all baseline methods. All required benchmarks are either public datasets that are automatically downloaded by the scripts from their official sources, or small data files included directly in the repository. Please refer to `README.md` for further details.

Acknowledgements

This work was partially supported by the NSF (IIS-2214141, ITE-2452834), ARO (W911NF-25-1-0059), ONR (N000142412663), and Amazon.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.

[2] Wael Alghamdi, Shahab Asoodeh, Flavio P Calmon, Oliver Kosut, Lalitha Sankar, and Fei Wei. Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1838–1843, 2022.

[3] Borja Balle, Gilles Barthe, Marco Gaboardi, Justin Hsu, and Tetsuya Sato. Hypothesis testing interpretations and renyi differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 2496–2506. PMLR, 2020.

[4] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.

[5] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[6] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[7] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.

[8] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.

[9] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.

[10] Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical measurement of information leakage. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 390–404. Springer, 2010.

[11] Xiao Chen, Peter Kairouz, and Ram Rajagopal. Understanding compressive adversarial privacy. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6824–6831. IEEE, 2018.

[12] Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.

[13] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54, 2016.

[14] David Dittrich, Erin Kenneally, et al. The menlo report: Ethical principles guiding information and communication technology research. Technical report, US Department of Homeland Security, 2012.

[15] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47, 1975.

[16] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1401–1408. IEEE, 2012.

[17] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12, 2006.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.

[19] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

[20] Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2017.

[21] Ronald Aylmer Fisher. Iris. uci machine learning repository. *DOI: https://doi. org/10.24432/C56C76*, 1988.

[22] Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 89–99. PMLR, 2020.

[23] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360, 2009.

[24] Ziv Goldfeld and Kristjan Greenewald. Sliced mutual information: A scalable measure of statistical dependence. *Advances in Neural Information Processing Systems*, 34:17567–17578, 2021.

[25] Jasper Goseling and Milan Lopuhaä-Zwakenberg. Robust optimization for local differential privacy. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1629–1634. IEEE, 2022.

[26] Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.

[27] Mangesh Gupte and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 135–146, 2010.

[28] Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pages 760–770. PMLR, 2021.

[29] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Generative adversarial privacy. *arXiv preprint arXiv:1807.05306*, 2018.

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

[31] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2019.

[32] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.

[33] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.

[34] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM Transactions on Database Systems (TODS)*, 39(1):1–36, 2014.

[35] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[36] Guy Lebanon, Monica Scannapieco, Mohamed Fouad, and Elisa Bertino. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. *Transactions on Data Privacy*, 2009.

[37] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

[38] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

[39] Milan Lopuhaä-Zwakenberg and Jasper Goseling. Mechanisms for robust local differential privacy. *Entropy*, 26(3):233, 2024.

[40] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.

[41] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2015.

[42] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM*

SIGSAC conference on computer and communications security, pages 634–646, 2018.

[43] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.

[44] Daniel P Palomar and Sergio Verdú. Gradient of mutual information in linear vector gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, 2005.

[45] Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe. On the equivalence between stein and de bruijn identities. *IEEE Transactions on Information Theory*, 58(12):7045–7067, 2012.

[46] Clément Pierquin, Aurélien Bellet, Marc Tommasi, and Matthieu Boussard. Rényi pufferfish privacy: General additive noise mechanisms and privacy amplification by iteration via shift reduction lemmas. In *International Conference on Machine Learning (ICML 2024)*, 2024.

[47] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International conference on scale space and variational methods in computer vision*, pages 435–446. Springer, 2011.

[48] Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Pointwise maximal leakage. *IEEE Transactions on Information Theory*, 69(12):8054–8080, 2023.

[49] Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.

[50] Aras Selvi, Huikang Liu, and Wolfram Wiesemann. Differential privacy via distributionally robust optimization. *Operations Research*, 2025.

[51] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[52] Shuang Song, Yizhen Wang, and Kamalika Chaudhuri. Pufferfish privacy mechanisms for correlated data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1291–1306, 2017.

[53] Mayuri Sridhar, Hanshen Xiao, and Srinivas Devadas. Pac-private algorithms. *Cryptology ePrint Archive*, 2024.

[54] Aart J Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.

[55] Michel Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.

[56] Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pages 611–644. Springer, 2023.

[57] Hanshen Xiao and Srinivas Devadas. Pac privacy and black-box privatization. *IEEE Security & Privacy*, 23(4):92–97, 2025.

[58] Xiaokui Xiao and Yufei Tao. Output perturbation with query relaxation. *Proceedings of the VLDB Endowment*, 1(1):857–869, 2008.

[59] Tao Zhang, Bradley A Malin, Netanel Raviv, and Yevgeniy Vorobeychik. Differential confounding privacy and inverse composition. In *2025 IEEE International Symposium on Information Theory (ISIT)*, pages 1–6. IEEE, 2025.

[60] Tao Zhang and Yevgeniy Vorobeychik. Breaking the gaussian barrier: Residual-pac privacy for automatic privatization. *arXiv preprint arXiv:2506.06530*, 2025.

[61] Tao Zhang and Yevgeniy Vorobeychik. Sliced rényi pufferfish privacy: Directional additive noise mechanism and private learning with gradient clipping. *arXiv preprint arXiv:2512.01115*, 2025.

[62] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

A Discussion: PAC/R-PAC Privacy vs. Differential Privacy

In this section, we discuss the difference and the relationship between PAC/R-PAC Privacy and DP (Definition 5).

DP and PAC (and R-PAC) Privacy use different semantics and different privacy quantification metrics. DP considers the presence or absence of individual records as secrets and ensures that, regardless of external knowledge, an adversary with access to mechanism outputs makes similar conclusions whether or not any individual data record is included in the dataset [17]. DP employs the worst-case probabilistic *input-independent indistinguishability* to quantify the privacy risk, by using the max-divergence for pure DP and the hockey stick divergence for the approximated DP.

Unlike DP, PAC Privacy can protect secrets that go beyond individual data records. PAC Privacy measures privacy in terms of the adversary's difficulty in achieving accurate reconstruction, capturing the semantics of the *impossibility of customized adversarial inference* [57], where the adversary is assumed to be computationally-unbounded. PAC Advantage Privacy uses the posterior advantage Δ_f^δ (Definition 3) to quantify privacy risk, where Δ_f^δ depends on a chosen f-divergence, secret (e.g., data) entropy determined by $\mathcal{D}$, an attack model in terms of ρ. When f-divergence is instantiated as KL divergence, Δ_f^δ is upper bounded by mutual information that is uniform over all adversaries and admissible ρ.

R-PAC and PAC Privacy are two sides of the same coin: PAC quantifies leakage (e.g., via Δ_f^δ), while R-PAC quantifies the remaining privacy (e.g., via $\mathtt{R}_f^\delta$), linked exactly by $\mathtt{IntP}_f(\mathcal{D}) = \mathtt{R}_f^\delta + \Delta_f^\delta$ (equation (9)). R-PAC Privacy uses the same semantics as PAC Privacy uses the posterior disadvantage $\mathtt{R}_f^\delta$ (Definition 9) to quantify the remaining privacy after leakage. When KL divergence is used, the posterior disadvantage $\mathtt{R}_f^\delta$ is lower bounded by the conditional entropy, which is uniform over all adversaries and admissible ρ.

PAC (and R-PAC) Privacy provides a more general framework that quantifies the reconstruction hardness for any sensitive information that an adversary might seek to infer. This encompasses not only individual membership inference (a special case), but also broader privacy concerns such as data reconstruction within specified error bounds, identification of multiple participants, or recovery of sensitive attributes. Crucially, PAC Privacy operates under distributional assumptions about the data (or general secrets) generation process $\mathcal{D}$, enabling instance-based analysis that can potentially require less noise than worst-case DP guarantees. However, the automatic privatization procedures (e.g., Auto-PAC and Efficient-PAC) proposed to realize PAC Privacy certify the privacy guarantee via an MI budget β, enforcing Gaussian surrogate bound $\mathtt{LogDet}(\mathcal{M}(X), B) \le \beta$ as a sufficient condition.

Consequently, the delivered mechanisms inherit MI–based properties and caveats (e.g., data processing, distributional/average–case nature, standard composition scaling, and lack of worst–case indistinguishability unless additional constraints are imposed), even though the abstract PAC Privacy notion itself does not rely on MI. Our SR-PAC follows the MI principle but implements an MI bound that is tighter than the Gaussian surrogate bound $\mathtt{LogDet}(\mathcal{M}(X), B)$.

While PAC (resp. R-PAC) privacy uses the semantics of impossibility of customized adversarial inference and is independent of mutual information (resp. conditional entropy), Auto-PAC uses mutual information (resp. conditional entropy) to quantify privacy risk. In this section, we discuss the difference between mutual information (MI) as privacy quantification and the input-independent indistinguishability of DP.

What each notion protects. Now, we discuss what each privacy notion protects. Let $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ be a randomized mechanism. DP, independent of input distribution, protects *worst-case, per-individual input-independen indistinguishability* by ensuring a uniform bound $\ell(x,y) \le \varepsilon$ almost surely (up to δ in the (ε,δ) case), where $\ell(x,y) = \log \frac{P_{X|Y}(x|y)}{P_X(x)}$ is the privacy-loss random variable. However, DP does not, in general, ensure that the *average* leakage $\mathtt{MI}(X;Y)$ is small—indeed, $\mathtt{MI}(X;Y)$ can scale with the dataset size unless ε shrinks appropriately. In contrast, MI-based privacy constrains the *average information* leaked from inputs to outputs under a specific input distribution $\mathcal{D} \in \Delta(\mathcal{X})$. MI controls average leakage: $\mathtt{MI}(X;Y) = \mathbb{E}_{P_{XY}}[\ell(X,Y)] \le \beta$ upper-bounds the expected log-likelihood gain of an optimal Bayesian adversary. However, MI *does not* by itself bound the worst-case leakage $L \equiv \operatorname{ess\,sup} \ell$; in particular, $\mathtt{MI}(X;Y) \le \beta$ is compatible with $L = \infty$ (rare but arbitrarily large disclosures).

Worst-case vs. average-case guarantees. DP is a distribution-free, worst-case guarantee that must hold for all neighboring datasets and all adversaries, independent of any input distribution. By contrast, MI-based privacy is *distributional*: it controls *expected* leakage under an input distribution P_X, typically via $\mathtt{MI}(X;Y) \le \beta$ for $Y = \mathcal{M}(X)$. Because $\mathtt{MI}(X;Y) = \mathbb{E}_{P_{XY}}[\ell(X,Y)]$, the noise needed to enforce $\mathtt{MI}(X;Y) \le \beta$ depends on P_X: when most probability mass lies on inputs for which $\ell(X,Y)$ is typically small, less perturbation can suffice, and the resulting noise shape may be tailored to that distribution. At the same time, an MI budget does not by itself preclude rare high-leakage cases: if there exists a measurable event $E \subseteq \mathcal{X} \times \mathcal{Y}$ with $P_{XY}(E) = p$ and $\ell(x,y) \ge L$ for all $(x,y) \in E$, then $\mathtt{MI}(X;Y) \ge pL$; hence the constraint $\mathtt{MI}(X;Y) \le \beta$ forbids such a case only when $pL > \beta$ (and any "perfect disclosure" with $L = \infty$ is incompatible for all $p > 0$).

Name-and-shame example. One example of "rare high–leakage cases" is the *name-and-shame*. Let $E = \{(x,y) : y = x\}$ denote the event in which the mechanism reveals the input directly, occurring with probability p. On E, the per–sample leakage is $\ell(x,y) = \log \frac{p_{X|Y}(x|x)}{p_X(x)} = -\log p_X(x)$, which can be very large (and unbounded when p_X has heavy tails or continuous support). Thus this is a small–probability, high–leakage branch. In the discrete case with finite support, one has $\mathtt{MI}(X;Y) = p\mathcal{H}(X)$. Choosing $p = \beta/\mathcal{H}(X)$ makes $\mathtt{MI}(X;Y) = \beta$, which saturates the heuristic $I\mathtt{MI}(X;Y) \ge pL$ when L is interpreted as the average leakage $\mathcal{H}(X)$ on E. If one insists on the pointwise form from the paragraph, taking $L_0 = \operatorname{ess\,inf}_x(-\log p_X(x))$ yields $\mathtt{MI}(X;Y)(X;Y) \ge pL_0$, which still places the example in the same regime. Finally, if "name–and–shame" is modeled as perfect disclosure with continuous X, then $\ell = \infty$ on E and the constraint rules it out immediately, since $L = \infty$ is incompatible with any $p > 0$.

DP perspective on the name-and-shame example. To see why this example fundamentally conflicts with the DP notion of rare-but-exact disclosure, consider the per–record

"name–and–shame" mechanism M that, independently for each index i, outputs (i, x_i) with probability p and $\perp$ otherwise. Let x and x' be neighboring databases that differ only in record i, and define the event $E = (i, x_i)$. Then

$$\Pr[M(x) \in E] = p, \quad \Pr[M(x') \in E] = 0.$$

The $(\varepsilon, \bar{\delta})$–DP inequality for E reads $p \le e^{\varepsilon} \cdot 0 + \bar{\delta} = \bar{\delta}$, hence any $(\varepsilon, \bar{\delta})$ satisfied by M must obey $\bar{\delta} \ge p$. In particular, with the standard regime $\bar{\delta} \ll 1/n$ (negligible failure probability), such a mechanism is *not* DP for any finite ε; conversely, allowing $\bar{\delta} \ge p$ makes the guarantee vacuous on the p–fraction of runs that reveal (i, x_i) exactly.

A.1 Fair Comparison Under MIA

PAC Privacy and R-PAC Privacy (and also MI-based privacy) address complementary notions of privacy to DP. Neither framework dominates the other. To perform a fair comparison, we focus on the cases when the privacy budgets of DP and PAC/R-PAC Privacy are "equalized". In particular, we consider Membership Inference Attack (MIA) defined by Definition 10 in Appendix C in [60].

DP can be understood through the lens of membership inference success rates. Consider the membership inference scenario from Definition 10 (Appendix C in [60]), where we have a dataset of size $n = \frac{N}{2}$ (i.e., each individual data record has a 50% probability of being included in the selected subset X). If a mechanism $\mathcal{M}$ is $(\varepsilon, \bar{\delta})$-DP, then by [30, 33], an adversary's ability to successfully infer whether a specific individual record i is included in the dataset (i.e., posterior success rate $p_o = 1 - \delta_i$) is fundamentally limited:

$$p_o \le 1 - \frac{1 - \bar{\delta}}{1 + e^{\varepsilon}}. \tag{15}$$

This bound demonstrates how DP parameters directly translate into concrete limits on an adversary's inference capabilities in MIA. Thus, the maximal posterior success rate permitted by $(\varepsilon, \bar{\delta})$-DP is $1 - \frac{1-\bar{\delta}}{1+e^{\varepsilon}}$.

In addition, there is a relationship between the posterior success rate p_o and the mutual information [53] (derived from (1)):

$$p_o \log \frac{p_o}{\bar{p}} + (1 - p_o) \log \frac{1 - p_o}{1 - \bar{p}} \le \mathtt{MI}(X; \mathcal{M}(X)), \tag{16}$$

where $\bar{p}$ is the optimal prior success rate, which is $\max(r, 1 - r)$ with r as the subsampling rate that selects the dataset from a data pool. Thus, given a privacy budget $\mathtt{MI}(X; \mathcal{M}(X)) = \beta$ and a prior success rate $\bar{p}$, we can calculate the posterior success level p_o and, by (15), pin down ε for a chosen $\bar{\delta}$ so that DP has an "equivalent" budget to PAC. The corresponding R-PAC budget is $\mathcal{H}(X) - \beta$.

For per-individual membership, the relevant secret is the membership indicator for person i, and the mechanism output is $Y = \mathcal{M}(X)$. Let $U_i \in \{0, 1\}$ denote the membership indicator as specified in Appendix C of [60]. Since $U_i \to X \to Y$ forms a Markov chain, the data-processing inequality gives $\mathtt{MI}(U_i; Y) \le \mathtt{MI}(X; Y) = \beta$. The Bernoulli–KL inequality used above applies equally with $\mathtt{MI}(U_i; Y)$ on the right-hand side; replacing it by $\mathtt{MI}(X; Y)$ is therefore conservative and still yields a valid upper bound on the Bayes-optimal membership posterior success p_o. This validates using $\mathtt{MI}(X; Y)$ to compute $p_o(\beta, \bar{p})$ for MIA and then selecting $(\varepsilon, \bar{\delta})$ so that (15) enforces the same p_o for a fair, like-for-like comparison between DP and PAC/R-PAC.

A.2 Noise Magnitude

In this section, we discuss how they differ in *noise magnitude* under an *equalized privacy budget*. Concretely, we fix a mutual-information budget β for PAC/R-PAC; when contrasting with DP, we use the $(\varepsilon, \bar{\delta})$ that induces the *same* posterior-success level via the MI↔DP conversion described in Section A.1. Even at this matched budget, the required noise can vary substantially. We measure it by the total noise magnitude $V(\beta) \equiv \mathbb{E}\|B\|_2^2$ for outputs $Y = \mathcal{M}(X) + B$. Let the centered output covariance have eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p > 0$ on its informative p-dimensional subspace ($p \le d$), and write $R = \max_j \lambda_j$. We first present the *ideal* Auto-PAC baseline derived from the log-det MI bound, then the (SR-PAC) optimizer that tightens noise under the same β, and finally contrast both with classical DP mechanisms that must mask worst-case sensitivity in d dimensions. (Throughout, Auto-PAC refers to this ideal log-det calibration; the practical Algorithm 1 uses estimated eigenvalues and a stabilization $10cv/\beta$, yielding total noise magnitude $(\sum_j \sqrt{\hat{\lambda}_j + 10cv/\beta})^2/(2v)$, a conservative upper envelope of the ideal baseline.)

Auto-PAC. Let the (centered) mechanism output have covariance eigenvalues $\lambda_1 \ge \cdots \ge \lambda_p > 0$ (in its informative p-dimensional subspace). Auto-PAC calibrates *Gaussian* noise $B \sim \mathcal{N}(0, \Sigma_B)$ under an MI budget β, yielding the total noise magnitude

$$V_{\mathrm{PAC}}(\beta) \;=\; \mathbb{E}\|B\|_2^2 \;=\; \frac{\left(\sum_{j=1}^{p} \sqrt{\lambda_j}\right)^2}{2\beta}.$$

(When the exhibited calibration targets $\mathtt{MI} \le \frac{1}{2}$, this specializes to $V_{\mathrm{PAC}} = (\sum_j \sqrt{\lambda_j})^2$.) A general bound is

$$\left(\sum_{j=1}^{p} \sqrt{\lambda_j}\right)^2 \;\le\; p \sum_{j=1}^{p} \lambda_j \;\le\; p^2 R,$$

where $R \equiv \max_j \lambda_j$. Hence $V_{\mathrm{PAC}}(\beta) = O(p^2 R/\beta)$ in the worst case (and improves to $O(pR/\beta)$ if $\sum_j \lambda_j = O(R)$). *(In practice, Algorithm 1 uses estimated eigenvalues and a stabilization $10cv/\beta$, yielding total noise magnitude $(\sum_j \sqrt{\hat{\lambda}_j + 10cv/\beta})^2/(2v)$, which is a conservative upper envelope of the ideal log-det calibration.)* Because differential

privacy (DP) must mask *worst-case* changes in all d coordinates, the required noise for d-dimensional outputs typically grows like $\sqrt{d}$ (e.g., $O(\sqrt{d}/n)$ for mean queries with dataset size n)—the classic "curse of dimensionality." Thus, when the data are effectively low-rank ($p \ll d$), Auto-PAC already mitigates this dimensional blow-up.

SR-PAC. SR-PAC *optimizes* the full noise distribution under the same MI budget β and strictly improves (or matches) the Gaussian baseline:

- *Universal gain (non-Gaussian outputs):* For any non-Gaussian output $Z = \mathcal{M}(X)$, SR-PAC achieves

$$\mathbb{E}\|B_{\mathrm{SR}}\|_2^2 \;<\; \mathbb{E}\|B_{\mathrm{PAC}}\|_2^2 \quad \text{at the same } \beta,$$

  closing the conservativeness of Auto-PAC. (If Z is exactly Gaussian, the gap can vanish.)

- *Anisotropic allocation:* The Stackelberg-optimal covariance is provably anisotropic; variance is shifted toward directions with high leakage and away from benign ones, improving utility without violating the MI budget.

- *Zero-noise subspaces:* Under a mild separation of directional sensitivities, there exists a threshold β_{lab} such that for all $\beta \le \beta_{\mathrm{lab}}$ SR-PAC injects *no* noise on an s-dimensional task-critical subspace (e.g., the $k-1$ label directions in classification), reducing the order from $O(p)$ to $O(p-s)$ in those regimes.

Comparison to DP. Since SR-PAC pointwise dominates Auto-PAC for every β and Auto-PAC already avoids DP's $\sqrt{d}$-type growth, SR-PAC inherits—and sharpens—the dimensional advantage. Writing

$$V_{\mathrm{SR}}(\beta) \;=\; V_{\mathrm{PAC}}(\beta) \;-\; \Delta(\beta), \qquad 0 \le \Delta(\beta) \le V_{\mathrm{PAC}}(\beta),$$

we have $\Delta(\beta) > 0$ whenever Z is non-Gaussian. In high-dimensional tasks with modest informative rank p and harmless directions ($s > 0$), SR-PAC reduces noise from $O(p)$ down to $O(p-s)$ (at fixed β), yielding a strictly better privacy–utility trade-off than both Auto-PAC and classical DP.

A.3 Computational Complexity

In this section, we concisely characterize the computational complexity of $(1-\gamma)$-Confidence Auto-PAC (i.e., Algorithm 1) and SR-PAC. For simplicity, we still use Auto-PAC to refer to Algorithm 1. Let d be the dimension of the mechanism output $\mathcal{M}(X) \in \mathbb{R}^d$.

Auto-PAC. Let m be the number of Monte Carlo simulations (samples) used by Auto-PAC. In addition, let $C_{\mathcal{M}}$ denote the cost of one evaluation of the (black-box) mechanism $\mathcal{M}(\cdot)$. Auto-PAC first draws m i.i.d. samples $X^{(1)}, X^{(2)}, \ldots, X^{(m)}$. For each sample $X^{(k)}$, Auto-PAC then evaluates $y^{(k)} = \mathcal{M}(X^{(k)})$; let $O(mC_{\mathcal{M}})$ denote the corresponding costs. It then forms the empirical mean and full empirical covariance $\hat{\Sigma} \in \mathbb{R}^{d\times d}$, which costs $O(md^2)$ time and $O(d^2)$ memory. Finally, it performs an SVD/eigendecomposition of $\hat{\Sigma}$ and constructs Σ_B, which costs $O(d^3)$ time (full decomposition) and $O(d^2)$ memory. Overall, the cost of Auto-PAC is

- Time: $O(mC_{\mathcal{M}} + md^2 + d^3)$;
- Memory: $O(d^2)$.

Here, the d^3 SVD/eigendecomposition step dominates at large output dimension.

SR-PAC (Monte Carlo Stackelberg optimization). SR-PAC uses the same black-box sampling access to $\mathcal{M}(\cdot)$ but avoids $d \times d$ SVD operations. In Algorithm 2, each update uses a fresh Monte-Carlo batch of size m (lines 6–8 and 12–15). Let $C_{\mathcal{M}}$ be the cost of one evaluation of $\mathcal{M}(\cdot)$, let C_π be the cost of one forward/backward pass of the decoder, and let C_g be the cost of sampling and differentiating through the perturbation rule. Then each decoder-gradient step costs $O(m(C_{\mathcal{M}} + C_\pi + C_g))$ time (lines 6–9), and each leader update costs $O(m(C_{\mathcal{M}} + C_\pi + C_g))$ time (lines 12–15). Over T_λ leader iterations with decoder-update phases triggered every T_ϕ iterations (line 3), the total runtime is

$$O\big(T_\lambda\, m\,(C_{\mathcal{M}} + C_\pi + C_g) \;+\; N_{\mathrm{dec}} \cdot m\,(C_{\mathcal{M}} + C_\pi + C_g)\big),$$

where N_{dec} is the total number of decoder-gradient steps (e.g., $N_{\mathrm{dec}} \approx \lfloor T_\lambda / T_\phi \rfloor \cdot T_\phi$ in Algorithm 2). The memory cost is dominated by storing model parameters and one mini-batch:

$$O(p_\pi + p_g + md),$$

where p_π and p_g are the parameter counts of the decoder and perturbation rule, respectively, and md accounts for holding a batch of m outputs in $\mathbb{R}^d$ during a step. In particular, SR-PAC avoids the $O(d^2)$ memory footprint and $O(d^3)$ matrix-decomposition bottleneck of Auto-PAC.

DP. Sensitivity (i.e., the maximal possible change on the output when a single data record changes) is the key component of DP privatization via noise perturbation. However, computing sensitivity is, in general, NP-hard [58]. DP-SGD [1] is a *decompose-then-compose* privatization scheme for DP, which avoids explicit sensitivity computation. However, DP-SGD adds per-iteration per-example gradient clipping and noise, inducing utility drop and computational overhead in large-scale applications [38]. Since DP and PAC/R-PAC Privacy adopt fundamentally different semantics and different privatization schemes, it is not self-evident how to compare them fairly in terms of computational complexity.

Scalable sliced variants. When d is large, the SR-PAC principle also applies to *sliced* objectives (e.g., sliced conditional entropy via sliced mutual information; see Appendix H of [60]), which replaces high-dimensional estimation with r one-dimensional projections. This yields an additional factor $O(r)$ over sampling while avoiding $O(d^3)$ operations, i.e., $O(mC_{\mathcal{M}} + rmd)$ for estimating sliced leakage terms (plus the decoder-training term if used).