Breaking the Gaussian Barrier: Residual-PAC Privacy for Automatic Privatization

Tao Zhang
Computer Science & Engineering
Washington University in St. Louis
St. Louis, Missouri 63130
Email: tz636@nyu.edu

Yevgeniy Vorobeychik
Computer Science & Engineering
Washington University in St. Louis
St. Louis, Missouri 63130
Email: yvorobeychik@wustl.edu

***Abstract*—The Probably Approximately Correct (PAC) Privacy framework [1] provides a powerful instance-based methodology for certifying privacy in complex data-driven systems. However, existing PAC Privacy algorithms rely on a Gaussian mutual information upper bound. We show that this is in general too conservative: the upper bound obtained by these algorithms is tight if and only if the perturbed mechanism output is jointly Gaussian with independent Gaussian noise. To address the inefficiency inherent in the Gaussian-based approach, we introduce *Residual-PAC (RPAC) Privacy*, an f-divergence-based measure that quantifies the privacy remaining after adversarial inference. When instantiated with Kullback–Leibler divergence, Residual-PAC Privacy is governed by conditional entropy. Moreover, we propose Stackelberg Residual-PAC (SR-PAC) privatization mechanisms for RPAC Privacy, a game-theoretic framework that selects optimal noise distributions through convex bilevel optimization. Our approach achieves tight privacy budget utilization for arbitrary data distributions. Moreover, it naturally composes under repeated mechanisms and provides provable privacy guarantees with higher statistical efficiency. Numerical experiments demonstrate that SR-PAC certifies the target privacy budget while consistently improving utility compared to existing methods.**

1. Introduction

Artificial intelligence systems now operate across increasingly critical domains—from healthcare diagnostics and autonomous transportation to personalized financial services and national security infrastructure. While demonstrating powerful capabilities, these data-driven systems introduce significant privacy risks as their training and operational data becomes increasingly sensitive and complex. Individual information can be inadvertently exposed through seemingly harmless outputs. Given the expanding scale of modern data pipelines, robust and reliable privacy guarantees have become essential rather than merely desirable.

The past two decades have witnessed the emergence of numerous privacy definitions and frameworks designed to address information leakage risks. Among these, Differential Privacy (DP) [2] stands as perhaps the most influential, providing strong worst-case guarantees against adversarial inference. In its canonical formulation, DP quantifies the maximum change in output probabilities induced by modifying a single data point, thereby ensuring indistinguishability between neighboring datasets. This theoretical foundation has inspired substantial research and practical deployment in tools used by major organizations (e.g., Apple [3]) and governments (e.g., US Census Bureau [4]). Other privacy notions, such as Maximal Leakage [5], [6], mutual information-based criteria, such as Mutual Information Differential Privacy (MI-DP) [7], and Fisher information-based metric [8], [9], [10], have been proposed to capture different adversarial models and to offer alternative trade-offs between privacy and utility.

Provable privacy guarantees for modern data-processing algorithms face two major obstacles. First, worst-case frameworks like DP require computing global sensitivity, which is generally intractable (NP-hard [11]). In addition, computing the optimal privacy bound of DP under composition is, in general, a #P-complete under composition task [12]. Likewise, enforcing MI-DP requires a white-box analysis of every possible input distribution, which is seldom available. 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. Empirical or simulation-based methods (e.g., testing resistance to membership inference [13]) address specific threats but lack rigorous, adversary-agnostic assurances. Bridging this gap requires a new, broadly applicable framework that can quantify and enforce privacy risk without relying on ad-hoc sensitivity analyses.

A promising alternative has recently emerged: the *Probably Approximately Correct (PAC) Privacy* framework [1]. Drawing inspiration from PAC learning theory [14], PAC Privacy fundamentally redefines the privacy objective as the information-theoretic hardness of reconstructing sensitive data given arbitrary information disclosure processing. PAC Privacy shifts the paradigm from indistinguishability to inference impossibility by characterizing the probability that any adversary—regardless of strategy or computational power—can accurately infer private data under a specified reconstruction criterion.

At the heart of PAC Privacy lies the concept of *PAC Advantage Privacy (PAC-AP)*, which extends the framework

to arbitrary adversarial inference tasks by quantifying the adversary's improvement in success probability through an f-divergence measure. When the f-divergence is instantiated as the Kullback–Leibler (KL) divergence, PAC-AP specializes to a mutual information-based criterion, providing a natural and information-theoretic characterization of privacy risk. A closed-form upper bound on mutual information—derived from the maximum entropy property of the Gaussian distribution—serves as the foundation for automatic PAC privatization algorithms, which operate via end-to-end black-box simulations. This approach circumvents the need for direct mutual information computation and enables practitioners to certify privacy risk with high statistical confidence, even when the internal structure of the data processing mechanism is unknown or intractable. In addition, PAC Privacy also enjoys elegant composition properties: when the composed mechanisms are independent, the total privacy loss can be bounded by the sum of the individual mutual information budgets.

However, the upper bound provided by automatic PAC Privacy algorithms (see Theorem 3 and Corollary 2 of [1], and Theorem 1 of [15]) is generally conservative. In particular, for a given privacy budget, the mutual information achieved under the noise distribution constructed by these algorithms is strictly less than the designated privacy budget unless the mechanism output is Gaussian and the noise is independent, zero-mean Gaussian. The privacy budget is exactly attained—i.e., the true mutual information matches the privacy bound—if and only if these Gaussian conditions hold. As a result, this conservativeness leads to a "waste" of privacy budget.

To maximize privacy budget efficiency, PAC Privacy defines optimal perturbation (Definition 9 of [1]) as the noise distribution that minimizes utility loss (e.g., ℓ_2 norm or noise power) while ensuring the mutual information remains within the privacy budget. Crucially, this mutual information constraint intricately links the noise distribution to both the data distribution and mechanism, often making the identification of optimal noise intractable, especially when the joint distribution lacks a closed-form or exhibits complex dependencies.

We address the limitations imposed by the conservativeness of automatic PAC Privacy algorithms. First, we introduce the notion of *Residual-PAC Privacy*, using f-divergence, to quantify the privacy that remains after information has been leaked by a data processing mechanism. While PAC-AP measures the amount of privacy loss incurred by a given mechanism, our Residual-PAC Privacy framework provides a complementary perspective by characterizing the remaining privacy guarantee that persists despite adversarial inference. When f-divergence is instantiated as KL divergence, our 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.

We formally characterize the conservativeness from the automatic PAC Privacy algorithms. The closed-form upper bound obtained by Theorem 3 of [1] forms the theoretical foundation of the automatic PAC Privacy algorithms. Algorithm 1 of [1] leverages this result to perform end-to-end automatic privatization, systematically constructing a Gaussian noise distribution that implements any designated privacy budget as an upper bound on mutual information. Importantly, we show that the Gaussian noise distribution constructed by Algorithm 1 of [1] is the unique solution to a constrained optimization problem that seeks to minimize the noise magnitude (specifically, the trace of the covariance matrix of the Gaussian noise) subject to the constraint that the resulting Gaussian mutual information does not exceed the prescribed privacy budget. Thus, the PAC private mechanism derived by Algorithm 1 of [1] not only enforces the desired privacy bound, but does so in an optimally noise-efficient manner under the Gaussian assumption.

Inspired by this observation, we propose a novel approach, termed *Stackelberg Residual-PAC (SR-PAC)*, for automating Residual-PAC privatization by formulating the problem as a Stackelberg game, without requiring any white-box characterization of the conditional entropy. In this framework, the leader selects a noise distribution to perturb the mechanism by minimizing the magnitude of the noise or perturbation. The follower then chooses a stochastic inference strategy to recover the sensitive data, seeking to minimize the expected log score function of the strategy, with the constraint that the expected log score equals the privacy budget. The privacy budget in Residual-PAC privacy directly translates to a mutual information bound via the simple relation: mutual information equals data entropy minus the privacy budget. When the entire probability space is considered, this bilevel optimization problem reduces to a convex program. We rigorously 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.

We demonstrate that, in general, the SR-PAC approach achieves more efficient privacy budget utilization than the automatic PAC Privacy algorithms, resulting in less conservative privacy guarantees for the same privacy budget, except in the special case where the overall distribution of the perturbed mechanism output is Gaussian, in which case both methods coincide and tightly implement the privacy budget. Besides inheriting PAC Privacy's additive composition property, SR PAC attains tighter privacy bounds by solving a convex optimization problem. Moreover, when the utility loss function is convex, our method admits an explicit convex optimization formulation, enabling the computation of optimal perturbations, which is a challenge that remained open in the original PAC Privacy framework. Finally, we validate these theoretical advances through numerical experiments, which consistently confirm the superior performance and practical advantages of SR-PAC over PAC Privacy.

1.1. Related Work

Privacy Quantification Notions. Quantitative notions of privacy leakage have been extensively studied across a variety of contexts, leading to mathematically rigorous frame-

works for assessing the amount of sensitive information that can be inferred by adversaries. Differential privacy (DP) and its variants have become the gold standard for formal privacy guarantees, with the original definitions by Dwork et al. [2], [16] formalizing privacy loss through bounds on the distinguishability of outputs under neighboring datasets. Variants such as concentrated differential privacy (CDP) [17], [18], zero-concentrated DP (zCDP) [19], and R'enyi differential privacy (RDP) [20] have further extended this framework by parameterizing privacy loss with different statistical divergences (e.g., R'enyi divergence), thereby enhancing flexibility in privacy accounting, especially for compositions and adaptive mechanisms. 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 [7], [21], with f-divergence and Fisher information offering finer-grained or context-specific metrics [1], [8], [9], [10]. 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, hypothesis testing privacy, 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 [22], [23], [24], [25], [26], [27], [28], [29], [30]. For example, Ghosh et al. [25] demonstrated that the geometric mechanism is universally optimal for differential privacy under certain loss-minimizing criteria in Bayesian settings, while Lebanon et al. [22] and Alghamdi et al. [29] studied utility-constrained optimization. Gupte et al. [26] 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 loss minimization.

Simulation-Based and Mutual Information Approaches Recent work has shifted toward instance-based and data-driven privacy frameworks that empirically measure information leakage for black-box mechanisms. PAC Privacy [1] and its improvements aim to automatically certify privacy loss by empirically estimating f-divergence and mutual information through repeated simulation and sampling. These approaches are especially valuable for complex mechanisms such as deep neural networks, where closed-form analysis or classical sensitivity calculations are infeasible or overly conservative. However, the mutual information constraint in PAC Privacy introduces substantial analytic and computational complexity, as it depends jointly on the data distribution, the mechanism, and the candidate noise distribution. In practice, the framework relies on propose-and-verify procedures, which may be conservative in the absence of tractable optimization.

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) [31] and related GAN-based frameworks [32], [33], [34]—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. [35] introduced a rigorous optimization framework for differential privacy 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 mechanism design.

2. Preliminaries: the PAC Privacy Framework

PAC privacy framework [1] considers the following general privacy problem. A sensitive input X is drawn from a distribution $\mathcal{D}$, which may be unknown or inaccessible. Each private data point x is defined over some measurable domain $\mathcal{X}^\dagger$, and the dataset $X \in \mathcal{X}$. There is a data processing (possibly randomized) mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y} \subset \mathbb{R}^d$, where $\mathcal{Y}$ is measurable. The central privacy concern is determining whether an adversary can accurately estimate the true input via an estimate $\widetilde{X}$ based on the observation $Y = \mathcal{M}(X)$, meeting some predefined success criterion captured by an indicator function ρ. PAC Privacy is formally defined as follows.

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

A user generates data X from distribution D *and sends $\mathcal{M}(X)$ to an adversary. The adversary who knows* 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. The function $\rho(\cdot, \cdot)$ 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 $|\widetilde{X} - X|_2 \leq \epsilon$ for some small $\epsilon > 0$; if X is a finite set of size n, success may be defined as correctly recovering more than $n - \epsilon$ elements. Notably, ρ need not admit a closed-form expression; it simply indicates whether the reconstruction satisfies the designated criterion for success.

This privacy definition is highly flexible by enabling ρ to encode a wide range of threat models and user-specified risk criteria. For example, in membership inference attacks [36], $\rho(\tilde{X}, X) = 1$ may indicate that $\widetilde{X}$ successfully determines the presence of a target data point u_0 in X. In reconstruction

attacks [37], 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:*

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

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

Definition 3 ($(\Delta_f^\delta, \rho, \mathcal{D})$ PAC Advantage Privacy [1])**.** *A mechanism $\mathcal{M}$ is termed $(\Delta_f\delta, \rho, \mathsf{D})$ PAC-advantage private if it is $(\delta, \rho, \mathsf{D})$ 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 $\mathcal{M}(X)$, measured as the *prior-expected posterior advantage* using f-divergence.

2.1. Automatic PAC Privatization Algorithms

PAC Privacy enables the automatic privatization for arbitrary black box mechanisms by bounding the *mutual information* between private data and the released output.

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

$$\mathtt{MI}(x; w) \equiv \mathcal{H}(x) - \mathcal{H}(x|w) = \mathcal{D}_{KL}(\mathsf{P}_{x,w}\|\mathsf{P}_x \otimes \mathsf{P}_w),$$

the KL-divergence between their joint distribution and the product of their marginals.

This approach supports simulation based privatization 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 [1] (hereafter "Auto-PAC") and the efficiency-improved version proposed in [15] (hereafter "Efficient-PAC").

Algorithm 1 Auto-PAC [1]

Require: deterministic mechanism M, dataset $\mathcal{D}$, sample size m, variance floor 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 $\hat{\Sigma} = \sum_{k=1}^m (y^{(k)} - \hat{\mu})(y^{(k)} - \hat{\mu})^T/m$.
Apply SVD: $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^T$, where $\hat{\Lambda}$ has eigenvalues $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \ldots \geq \hat{\lambda}_d$.
Find $j_0 = \arg\max_j \hat{\lambda}_j$ for $\hat{\lambda}_j > c$.
if $\min_{1\leq j\leq j_0, 1\leq l\leq d} |\hat{\lambda}_j - \hat{\lambda}_l| > r\sqrt{dc} + 2c$ **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 $\Sigma_{\boldsymbol{B}} = \hat{U}\Lambda_{\boldsymbol{B}}^{-1}\hat{U}^T$.
else
  Set $\Sigma_{\boldsymbol{B}} = (\sum_{j=1}^d \hat{\lambda}_j + dc)/(2v) \cdot \boldsymbol{I}_d$.
end if
Output: $\Sigma_{\boldsymbol{B}}$.

We start by introducing Auto-PAC. Consider a deterministic data processing mechanism $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$, where the output norm is uniformly bounded: $\|\mathcal{M}(X)\|_2 \leq r$ for all X. To ensure PAC Privacy, the mechanism is perturbed by Gaussian noise $B \sim \mathcal{N}(0, \Sigma_B)$, and the privacy leakage is quantified via mutual information between the input X and the noisy output $\mathcal{M}(X) + B$. For any deterministic mechanism $\mathcal{M}$ and noise B according to any distribution, define

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

where $\Sigma_{\mathcal{M}(X)}$ and Σ_B are the covariances of $\mathcal{M}(X)$ and B.

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

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

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 $\mathtt{MI}(X; \mathcal{M}(X) + B) \leq \frac{1}{2}$.

This result establishes a simple upper bound on the mutual information after perturbation and implies that the noise can be tailored anisotropically. To apply this bound in practice, the covariance $\Sigma_{\mathcal{M}(X)}$ must be estimated from Monte Carlo simulations. The authors of [1] propose a high-confidence noise calibration protocol (Algorithm 1) that determines an appropriate noise covariance Σ_B to ensure that $\mathtt{MI}(X; \mathcal{M}(X) + B) \leq v + \beta$ with confidence at least $1 - \gamma$, given user-specified parameters v, β, and sample complexity m. We refer to $\widehat{\beta} = v + \beta$ as the *privacy budget*.

Algorithm 2 Efficient-PAC [15]

Require: deterministic mechanism $\mathcal{M}$, data distribution $\mathcal{D}$, precision parameter τ, convergence function f_τ, privacy budget β, unitary projection matrix $A \in \mathbb{R}^{d\times d}$.

Initialize $m \leftarrow 1$, $\boldsymbol{\sigma}_0 \leftarrow$ null, $\mathbf{G} \leftarrow$ null
while $m \leq 2$ or $f_\tau(\boldsymbol{\sigma}_{m-1}, \boldsymbol{\sigma}_m) \geq \tau$ **do**
  Sample $X_m \sim \mathcal{D}$, compute $y_m \leftarrow \mathcal{M}(X_m)$
  Set $g_m \leftarrow [y_m \cdot A_1, \ldots, y_m \cdot A_d]$, append to $\mathbf{G}$
  Set $\boldsymbol{\sigma}_m[k]$ to empirical variance of column k in $\mathbf{G}$, increment $m \leftarrow m + 1$
end while
for $i = 1$ to d **do**
  Set $e_i \leftarrow \frac{\sqrt{\boldsymbol{\sigma}_m[i]}}{2\beta} \sum_{j=1}^{d} \sqrt{\boldsymbol{\sigma}_m[j]}$
end for
return Σ_B with $\Sigma_B[i][i] = e_i$

The framework extends naturally to *randomized mechanisms* of the form $\mathcal{M}(X, \theta)$, where θ is a random seed (Corollary 2 of [1]).

Recent work by Sridhar et al. [15] improves the practicality and efficiency of PAC Privacy (i.e., Efficient-PAC) by introducing an anisotropic noise calibration scheme that avoids full covariance estimation (Algorithm 2). Instead of computing the entire output covariance matrix and performing expensive matrix decomposition, their method projects mechanism outputs onto a unitary basis and estimates only the per-direction variances. This leads to a more scalable and sample-efficient algorithm while still ensuring rigorous mutual information guarantees under the PAC Privacy framework.

In addition, the authors address a key limitation of black-box privacy mechanisms—namely, the instability of outputs caused by random seeds, arbitrary encodings, or non-deterministic implementations. They propose methods for reducing such instability through output regularization and canonicalization, enabling more consistent noise calibration and better overall utility. These refinements are particularly impactful in high-dimensional or structure-sensitive learning tasks, where the original PAC scheme may incur unnecessary noise due to variability not intrinsic to the learning objective.

Theorem 2 (Theorem 1 of [15])**.** *Let $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$ be a deterministic mechanism, and let $A \in \mathbb{R}^{d\times d}$ be a unitary projection matrix. Let $\boldsymbol{\sigma} \in \mathbb{R}^d$ be the variance vector of the projected outputs $\mathcal{M}(X) \cdot A$, and let $B \sim \mathcal{N}(0, \Sigma_B)$ be the additive noise with covariance $\Sigma_B = \mathrm{diag}(e_1, \ldots, e_d)$, where $e_i = \frac{\sqrt{\sigma_i}}{2\beta} \sum_{j=1}^{d} \sqrt{\sigma_j}$. Then, the mutual information between the input and privatized output satisfies $\mathtt{MI}(X; \mathcal{M}(X) + B) \leq \beta$.*

3. Residual-PAC Privacy

PAC Advantage Privacy, as defined in Definition 3, quantifies the amount of *privacy leakage*—denoted by Δ_f^δ—incurred from releasing $\mathcal{M}(X)$, measured as the (prior-expected)*posterior advantage*, using f-divergence, gained by the adversary after observing $\mathcal{M}(X)$. 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 information has been leaked through the mechanism's output. To formalize this residual protection, we first define the *intrinsic privacy* of a data distribution $\mathcal{D}$ as its closeness to the uniform distribution over $\mathcal{X}$. Specifically, let $\mathcal{U}$ denote the uniform distribution on $\mathcal{X}$, and define

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

where $\mathtt{D}_f(\mathcal{D}\|\mathcal{U})$ is the f-divergence between $\mathcal{D}$ and $\mathcal{U}$, quantifying how much $\mathcal{D}$ deviates from perfect uniformity. Intuitively, $-\mathtt{D}_f(\mathcal{D}\|\mathcal{U})$ serves as a reward for uniformity (i.e., maximal intrinsic privacy), and by construction, $\mathtt{IntP}_f(\mathcal{D}) \leq 0$, attaining zero only when $\mathcal{D}$ is uniform.

Definition 5 ($(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC Privacy)**.** *A mechanism $\mathcal{M}$ is said to be $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC private if it is $(\delta, \rho, \mathcal{D})$ PAC private and*

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

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

Under this definition, $\mathtt{R}_f^\delta$ characterizes the *residual privacy guarantee*—that is, the portion of intrinsic privacy that remains uncompromised after the adversary's inference. Notably, the total intrinsic privacy is precisely decomposed as

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

where Δ_f^δ is the PAC Advantage Privacy, or privacy leakage, term. This relationship provides a complete and interpretable accounting of privacy risk, distinguishing between the privacy that is lost and that which endures after information disclosure.

3.1. Foundation of Residual-PAC Privacy

In this section, we develop general results to support concrete analyses under the Residual-PAC Privacy framework. We begin by introducing key information-theoretic quantities, entropy and conditional entropy, which are fundamental to our development.

Definition 6 (Entropy and Conditional Entropy)**.** *Let X be a random variable on a discrete alphabet $\mathcal{X}$ with probability mass function $P_X(x) = \Pr(X = x)$. The* Shannon entropy *of X is*

$$\mathcal{H}(X) = -\sum\nolimits_{x\in\mathcal{X}} P_X(x)\, \log P_X(x).$$

If X instead has a continuous distribution on $\mathcal{X} \subseteq \mathbb{R}^n$ with probability density function $f_X(x)$, its differential entropy *is*

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

Moreover, let (X, W) be jointly distributed. If X, W are discrete with joint PMF $p_{X,W}(x, w)$ and marginals $p_W(w)$, the conditional entropy *of X given W is*

$$\mathcal{H}(X|W) = \sum\nolimits_{w\in\mathcal{W}} P_W(w) \Big[- \sum\nolimits_{x\in\mathcal{X}} P_{X|W}(x \mid w) \log P_{X|W}(x|w) \Big].$$

If (X, W) are continuous with joint PDF $f_{X,W}(x, w)$ and marginal $f_W(w)$, the differential conditional entropy *is*

$$h(X|W) = \int_{\mathcal{W}} f_W(w) \Big[- \int_{\mathcal{X}} f_{X|W}(x|w) \log f_{X|W}(x|w)\, dx \Big] dw.$$

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.

Mutual information plays a central role in the PAC Privacy analysis, as it is defined in terms of entropy and conditional entropy. Let X and W be random variables with joint distribution $P_{X,W}(x, w)$ and marginals $P_X(x)$ and $P_W(w)$. Then, the mutual information between X and W is given by

$$\mathtt{MI}(X; W) = \mathcal{H}(X) - \mathcal{H}(X|W) = \mathtt{D}_{\mathrm{KL}}\big(P_{X,W} \,\|\, P_X \otimes P_W\big),$$

where $\mathtt{D}_{\mathrm{KL}}\big(P_{X,W} \,\|\, P_X \otimes P_W\big)$ denotes the Kullback–Leibler (KL) divergence between the joint distribution $P_{X,W}$ and the product of the marginal distributions $P_X \otimes P_W$.

Theorem 1 of [1] lays the foundation of PAC Privacy analysis, which provides a quantitative upper bound on the adversary's inference advantage in terms of f-divergence. Specifically, for any f-divergence $\mathtt{D}_f$, the adversary's posterior advantage $\Delta_f\delta$ is bounded by the minimum f-divergence between the joint distribution $(X, \mathcal{M}(X))$ and the product of the marginal P_X and any auxiliary output distribution P_W that is independent of X:

$$\Delta_f\delta \le \inf_{P_W} \mathtt{D}_f\big(PX, \mathcal{M}(X) \,\|\, P_X \otimes P_W\big), \tag{2}$$

where $P_{X,\mathcal{M}(X)}$ denotes the joint distribution of the data and mechanism output, respectively, and P_W ranges over all distributions on the output space.

When the f-divergence is instantiated as the KL divergence and P_W is chosen as the marginal output distribution $P_{\mathcal{M}(X)}$, this bound becomes the mutual information between X and $\mathcal{M}(X)$:

$$\Delta_{\mathrm{KL}}\delta \le \mathtt{MI}(X; \mathcal{M}(X)),$$

showing that the mutual information between the input and the mechanism output governs the worst-case posterior advantage.

For any selected f-divergence $\mathtt{D}_f$, the inequality (2) implies that a mechanism $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC Privacy if

$$\mathtt{R}_f^\delta \ge \mathtt{IntP}_f(\mathcal{D}) - \inf_{P_W} \mathtt{D}_f\left(P_{X,\mathcal{M}(X)} \,\|\, P_X \otimes P_W\right). \tag{3}$$

Corollary 1. *Let $\mathtt{D}_f$ be the KL-divergence. $\mathcal{M} : \mathcal{X} \to \mathcal{Y}$ satisfies $(\mathtt{R}_f^\delta, \rho, \mathcal{D})$ Residual-PAC Privacy if*

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

where $\mathtt{V} = \log(|\mathcal{X}|)$ if $\mathcal{H}$ is Shannon entropy, and $\mathtt{V} = \log(\int_{\mathcal{X}} dx)$ if $\mathcal{H}$ is differential entropy.

Proof. By Theorem 1 of [1], a mechanism $\mathcal{M}$ satisfies $(\delta, \rho, \mathcal{D})$-PAC privacy where $\mathtt{D}_{\mathrm{KL}}(\mathbf{1}_\delta \| \mathbf{1}_{\delta_\rho^o}) \le \mathtt{MI}(X; \mathcal{M}(X))$. Thus, $\mathtt{R}_{\mathrm{KL}}^\delta \ge \mathtt{IntP}_{\mathrm{KL}}(\mathcal{D}) - \inf_{P_W} \mathtt{D}_{\mathrm{KL}}\left(P_{X,\mathcal{M}(X)} \,\|\, P_X \otimes P_W\right) \ge \mathtt{IntP}_{\mathrm{KL}}(\mathcal{D}) - \mathtt{MI}(X; \mathcal{M}(X))$, where $\mathtt{IntP}_{\mathrm{KL}}(\mathcal{D}) = -\mathtt{D}_{\mathrm{KL}}(\mathcal{D}\|\mathcal{U}) = \mathcal{H}(X) - \mathtt{V}$, where $\mathtt{V} = \log(|\mathcal{X}|)$ if $\mathcal{H}$ is Shannon entropy, and $\mathtt{V} = \log(\int_{\mathcal{X}} dx)$ if $\mathcal{H}$ is differential entropy. Thus, we get $\mathtt{R}_f^\delta \ge \mathcal{H}(X|\mathcal{M}(X)) - \mathtt{V}$. □

Corollary 1 follows Theorem 1 of [1] establishes that the residual privacy $\mathtt{R}_f^\delta$ is lower bounded by the conditional entropy $\mathcal{H}(X|\mathcal{M}(X))$ up to a constant $\mathtt{V}$, where $\mathtt{V}$ is independent of both the data distribution $\mathcal{D}$ and the mechanism $\mathcal{M}$. As such, $\mathtt{R}_f^\delta$ can, without loss of generality, be offset by $\mathtt{V}$ in Residual-PAC Privacy analysis (i.e., $\mathtt{R}_f^\delta - \mathtt{V}$), effectively serving as a privacy quantification that is lower-bounded by the conditional entropy $\mathcal{H}(X|\mathcal{M}(X))$.

4. Characterizing The Gaussian Barrier of Automatic PAC Privatization

In this section, we characterize the utility of Auto-PAC algorithms (Algorithm 1) with a focus on the conservativeness of the mutual information bounds they implement. The resulting perturbation bound is conservative due to a nonzero discrepancy between the true mutual information and the upper bound established by Theorem 1:

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

where $\mathtt{LogDet}(\mathcal{M}(X), B)$ is defined by (1). Here, the algorithm implements output perturbation of the form $\mathcal{M}(X) + B$, where $B \sim Q_B = \mathcal{N}(0, \Sigma_B)$ is Gaussian noise independent of the mechanism output $\mathcal{M}(X)$. Define $Z = \mathcal{M}(X) + B$. Then, Z has mean $\mu_Z = \mu_{\mathcal{M}(X)}$ and covariance $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$, where $\Sigma_{\mathcal{M}(X)}$ denotes the covariance of $\mathcal{M}(X)$. In addition, let $P_{\mathcal{M},B}$ denote the true distribution of $Z = \mathcal{M}(X) + B$. Let

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

denote the *Gaussian distribution with the same first and second moments as Z*.

Proposition 1. *Let $B \sim \mathcal{N}(0, \Sigma_B)$. Then, $\mathtt{Gap}_{\mathtt{d}} = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_{\mathcal{M}}) \ge 0$, where $\mathtt{Gap}_{\mathtt{d}} = 0$ if and only if $P_{\mathcal{M},B} = \mathcal{N}(\mu_Z, \Sigma_Z)$.*

Proof. Recall that $Z = \mathcal{M}(X) + B$. Since $Z \mid X \sim \mathcal{N}(\mathcal{M}(X), \Sigma_B)$, we have by the definition of mutual information: $\mathtt{MI}(X; Z) = \mathcal{H}(Z) - \mathcal{H}(Z \mid X) = \mathcal{H}(Z) - \mathcal{H}(B)$. Now consider the reference distribution $\widetilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$, where $\mu_Z = \mu_{\mathcal{M}(X)}$

and $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$. Its differential entropy is given by $\mathcal{H}(\widetilde{Q}_{\mathcal{M}}) = \frac{1}{2}\log\left[(2\pi e)^d \det(\Sigma_Z)\right]$, and similarly, $\mathcal{H}(B) = \frac{1}{2}\log\left[(2\pi e)^d \det(\Sigma_B)\right]$. Hence, $\frac{1}{2}\log\det\left(I_d + \Sigma_{\mathcal{M}(X)}\Sigma_B^{-1}\right) = \frac{1}{2}\log\left(\frac{\det(\Sigma_Z)}{\det(\Sigma_B)}\right) = \mathcal{H}(P_Z) - \mathcal{H}(B)$. Combining this with the earlier expression for $\mathtt{MI}(X;Z)$, we obtain $\mathtt{Gap}_{\mathrm{d}} = \Big[\mathcal{H}(\widetilde{Q}_{\mathcal{M}}) - \mathcal{H}(B)\Big] - [\mathcal{H}(Z) - \mathcal{H}(B)] = \mathcal{H}(\widetilde{Q}_{\mathcal{M}}) - \mathcal{H}(Z)$. By the definition of KL divergence, we have $\mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \,\|\, \widetilde{Q}_{\mathcal{M}}) = \mathcal{H}(\widetilde{Q}_{\mathcal{M}}) - \mathcal{H}(Z)$. Therefore, $\mathtt{Gap}_{\mathrm{d}} = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \,\|\, \widetilde{Q}_{\mathcal{M}}) \geq 0$, with equality if and only if $P_{\mathcal{M},B} = \widetilde{Q}_{\mathcal{M}}$, i.e., Z is exactly Gaussian with distribution $\mathcal{N}(\mu_Z, \Sigma_Z)$. □

Proposition 1 shows that the *Gaussianity* of $\mathcal{M}(X)$ in terms of $\mathtt{Gap}_{\mathrm{d}}$ is equivalent to the KL divergence $\mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B}\|\widetilde{Q}_{\mathcal{M}})$. Let $\widetilde{Z} \sim \widetilde{Q}_{\mathcal{M}}$. Thus, $\mathtt{MI}(X;\widetilde{Z}) = \mathtt{LogDet}(\mathcal{M}(X), B)$.

Proposition 2. *For any privacy budget $\beta > 0$, the noise distribution $Q_B^* = \mathcal{N}(0, \Sigma_B^*)$ obtained by Algorithm 1 is a unique solution of the following problem:*

$$\inf_{B\sim Q_B} \mathbb{E}_{Q_B}\left[\|B\|_2^2\right], \text{ s.t. } \mathtt{MI}(X;\widetilde{Z}) \leq \beta \text{ with } \widetilde{Z} \sim \widetilde{Q}_{\mathcal{M}}. \tag{5}$$

4.1. Mechanism Comparison in PAC Privacy

Definition 9 of [1] defines the optimal perturbation that tightly implements the privacy budget while maintaining optimal utility, where the utility is captured by some loss function $\mathcal{K}$. That is, an optimal perturbation, Q^* is a solution of the following optimization problem:

$$\begin{aligned} &\inf_Q \mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B;\mathcal{M})], \\ &\text{s.t. } \mathtt{MI}(X;\mathcal{M}(X)+B) \leq \beta, B \sim Q. \end{aligned} \tag{6}$$

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

Now, we show 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.

Proposition 3. *Fix a mechanism $\mathcal{M}$ and a data distribution $\mathcal{D}$. Let $\mathcal{Q}$ denote the collection of all zero-mean noise distributions under consideration, and let $\mathtt{I}_{\mathrm{true}} : \mathcal{Q} \mapsto \mathbb{R}_{\geq 0}$ be the true mutual information functional; i.e., $\mathtt{I}_{\mathrm{true}}(Q) = \mathtt{MI}(X;\mathcal{M}(X)+B)$ with $B \sim Q$ for $Q \in \mathcal{Q}$. For each privacy budget $\beta \geq 0$, we define the feasible region $\mathcal{F}(\beta) \equiv \{Q \in \mathcal{Q} : \mathtt{I}_{\mathrm{true}}(Q) \leq \beta\}$. 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 following problem:*

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

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

Proof. Since $\beta_1 < \beta_2$, any distribution Q satisfying $I_{\mathrm{true}}(Q) \leq \beta_1$ necessarily satisfies $I_{\mathrm{true}}(Q) \leq \beta_2$. Consequently, we have the inclusion $\mathcal{F}(\beta_1) \subseteq \mathcal{F}(\beta_2)$. Let A and $\hat{A}$ be arbitrary sets with $A \subseteq \hat{B}$, and let f be any real-valued function defined on B. Then, $\inf_{x\in A} f(x) \geq \inf_{x\in\hat{A}} f(x)$, with equality holding when the infimum over $\hat{A}$ is attained within the subset A. Applying this with $A = \mathcal{A}(\beta_1)$, $\hat{A} = \mathcal{F}(\beta_2)$, and $f(Q) = \mathbb{E}_Q\left[\|B\|_2^2\right]$ yields $\inf_{Q\in\mathcal{F}(\beta_1)} \mathbb{E}_Q\left[\|B\|_2^2\right] \geq \inf_{Q\in\mathcal{F}(\beta_2)} \mathbb{E}_Q\left[\|B\|_2^2\right]$. By definition, $Q^*(\beta_i)$ achieves the infimum of $\mathbb{E}_Q\left[\|B\|_2^2\right]$ over $\mathcal{F}(\beta_i)$ for $i = 1, 2$. Therefore, $\mathbb{E}_{Q^*(\beta_1)}\left[\|B\|_2^2\right] = \inf_{Q\in\mathcal{F}(\beta_1)} \mathbb{E}_Q\left[\|B\|_2^2\right] \geq \inf_{Q\in\mathcal{F}(\beta_2)} \mathbb{E}_Q\left[\|B\|_2^2\right] = \mathbb{E}_{Q^*(\beta_2)}\left[\|B\|_2^2\right]$. □

However, the automatic PAC privatization Algorithm 1 solving the optimization problem (5) considers the conservative implementation of a given privacy budget. Next result shows that when $\mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B}\|\widetilde{Q}_{\mathcal{M}}) > 0$, Algorithm 1 in general does not maintain coherent ordering PAC Privacy.

With a slight abuse of notation, for any mechanism $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$, let $\mathtt{Gap}_{\mathrm{d}}(Q) = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B}\|\widetilde{Q}_{\mathcal{M}})$, where $B \sim Q$.

Theorem 3. *Fix a mechanism $\mathcal{M}$ and a data distribution $\mathcal{D}$. Let $\mathcal{Q}$ denote the collection of all zero-mean noise distributions under consideration, and let $\mathtt{I}_{\mathrm{true}} : \mathcal{Q} \mapsto \mathbb{R}_{\geq 0}$ be the true mutual information functional; i.e., $\mathtt{I}_{\mathrm{true}}(Q) = \mathtt{MI}(X;\mathcal{M}(X)+B)$ 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*

$$\mathtt{G}(\beta_2, \beta_1) \equiv \mathtt{Gap}_{\mathrm{d}}(Q(\beta_2)) - \mathtt{Gap}_{\mathrm{d}}(Q(\beta_1)).$$

Then, we have:

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

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

Theorem 3 provides an exact characterization of how the *actual* information leakage $\mathtt{I}_{\mathrm{true}} = \beta - \mathtt{Gap}_{\mathrm{d}}$ behaves when one adjusts the nominal Gaussian-surrogate privacy budget β. Increasing the budget from β_1 to $\beta_2 > \beta_1$ "permits" an extra $\beta_2 - \beta_1$ of leakage under the surrogate bound, but part of that allowance may be "wasted" if the mechanism's output remains non-Gaussian. The wasted portion is

$$\mathtt{Gap}_{\mathrm{d}}\big(Q^*(\beta_2)\big) - \mathtt{Gap}_{\mathrm{d}}\big(Q^*(\beta_1)\big).$$

The theorem shows that if this wasted gap does not exceed $\beta_2 - \beta_1$, then I_{true} indeed increases (or stays constant) with the budget; if it exceeds $\beta_2 - \beta_1$, then paradoxically I_{true} decreases despite a larger nominal budget, because the Gaussian-surrogate bound becomes overly conservative.

In applications, this result cautions against comparing or calibrating privacy mechanisms solely by their surrogate budgets. Two mechanisms with identical β may exhibit very different true leakages if their Gaussianity gaps differ. To ensure that loosening the surrogate constraint actually increases real leakage (and therefore yields the expected utility–privacy trade-off), one should estimate or bound how

$\texttt{Gap}_\texttt{d}$ grows with β. If that gap grows too rapidly—so that the increase in conservativeness outstrips the nominal allowance—it may be necessary to incorporate direct estimates of true mutual information into the noise-calibration procedure rather than relying exclusively on the Gaussian bound.

5. $\texttt{Gap}_\texttt{d}$ Reduction via Non-Gaussianity Correction

In this section, we propose two approaches to reduce $\texttt{Gap}_\texttt{d}$ for a given Σ_B that is obtained by Algorithm 1, so that we can obtain a better estimation of the true mutual information under the perturbation using $B \sim \mathcal{N}(0, \Sigma_B)$.

For any deterministic $\mathcal{M}$ and a Gaussian noise $B \sim \mathcal{N}(0, \Sigma_B)$, recall the definition of $\widetilde{Q}_\mathcal{M} = \mathcal{N}(\mu_Z, \Sigma_Z)$ in (4), where $Z = \mathcal{M}(X) + B$ with mean μ_Z and covariance Σ_Z. In addition, let $\texttt{D}_Z = \texttt{D}_{\mathrm{KL}}(P_{\mathcal{M},B} \| \widetilde{Q}_\mathcal{M})$. By Proposition 1, we know $\texttt{Gap}_\texttt{d} = \texttt{D}_Z$. For any estimator $\widehat{\texttt{D}}_Z$ of $\texttt{D}_Z$, define

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

Then, for $0 < \widehat{\texttt{D}}_Z \leq \texttt{D}_Z$, we have

$$\texttt{MI}(X; \mathcal{M}(X) + B) \leq \texttt{IMI}(\widehat{\texttt{D}}_Z) < \texttt{LogDet}(\mathcal{M}(X), B).$$

That is, if we can obtain $\widehat{\texttt{D}}_Z$ satisfying $0 < \widehat{\texttt{D}}_Z \leq \texttt{D}_Z$, then for any Σ_B that guarantee $\texttt{LogDet}(\mathcal{M}(X), B) = \beta$, we have $\texttt{IMI}(\widehat{\texttt{D}}_Z) = \beta - \texttt{D}_Z$ as an estimation of the true mutual information.

Before describing the approaches, we first introduce the *denoising score matching (DSM)* [38]. DSM is a method to learn the score function $\nabla_z \log P_Z(z)$ of a random vector $Z \in \mathbb{R}^d$ without explicit density estimation.

Definition 7 (DSM-Optimal Score Estimator)**.** *Given i.i.d. samples $\{z_i\}_{i=1}^N$ from a distribution P_Z over $\mathbb{R}^d$:*

1) *Center data: $\widetilde{z}_i = z_i - \widehat{\mu}_Z$ where $\widehat{\mu}_Z = \frac{1}{N}\sum_{i=1}^N z_i$.*
2) *Perturb samples: $\widetilde{z}_i' = \widetilde{z}_i + \sqrt{\epsilon} v_i$ with $v_i \sim \mathcal{N}(0, I_d)$, $\epsilon > 0$.*
3) *Train parametric score model: Optimize α for $s_\alpha : \mathbb{R}^d \to \mathbb{R}^d$ via: $\alpha^* = \arg\min_\alpha \mathcal{L}_{\mathrm{DSM}}(\alpha)$, where*

$$\mathcal{L}_{\mathrm{DSM}}(\alpha) = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{v_i}\left[\left\| s_\alpha(\widetilde{z}_i') + \frac{v_i}{\sqrt{\epsilon}} \right\|_2^2\right].$$

The resulting function s_{α^*} is the DSM-optimal score estimator for the centered distribution $P_{\widetilde{Z}}$, satisfying:

$$s_{\alpha^*}(\widetilde{z}) \approx \nabla_{\widetilde{z}} \log P_{\widetilde{Z}}(\widetilde{z}),$$

with convergence $s_{\alpha^*} \to \nabla \log P_{\widetilde{Z}}$ as $\epsilon \to 0^+$ under mild conditions [38], [39].

Next, we provide two schemes to compute $\widehat{\texttt{D}}_Z$.

Theorem 4 (Stein-Discrepancy Bound)**.** *Let X be any random variable, $\mathcal{M} : \mathcal{X} \to \mathbb{R}^d$ ($d \geq 1$) a deterministic mechanism, and $B \sim \mathcal{N}(0, \Sigma_B)$ independent Gaussian noise. Define $Z = \mathcal{M}(X) + B$ with mean $\mu_Z = \mathbb{E}[Z]$ and covariance $\Sigma_Z = \mathrm{Cov}(Z)$. Let $\texttt{D}_Z = \texttt{D}_{\mathrm{KL}}(P_Z \,\|\, \mathcal{N}(\mu_Z, \Sigma_Z))$.*

Given i.i.d. samples $\{z_i\}_{i=1}^N$:

1) *Compute $\widehat{\mu}_Z = \frac{1}{N}\sum_{i=1}^N z_i$ and $\widetilde{z}_i = z_i - \widehat{\mu}_Z$;*
2) *Compute $\widehat{\Sigma}_Z = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_i \widetilde{z}_i^\top$;*
3) *Let s_{α^*} be the DSM-optimal score estimator (Definition 7) for $\{\widetilde{z}_i\}_{i=1}^N$.*

Define the estimator:

$$\widehat{\texttt{D}}_Z = \frac{1}{2N}\sum_{i=1}^N \|s_{\alpha^*}(\widetilde{z}_i)\|_2^2 - \frac{1}{N}\sum_{i=1}^N \nabla \cdot s_{\alpha^*}(\widetilde{z}_i),$$

where $\nabla \cdot s_{\alpha^*}(\widetilde{z}_i) = \sum_{k=1}^d \frac{\partial [s_{\alpha^*}]_k}{\partial z_k}(\widetilde{z}_i)$. Then:

$$\mathbb{E}\left[\widehat{\texttt{D}}_Z\right] \leq \frac{1}{2}\mathbb{E}\left[\|\nabla_z \ln p_Z(Z)\|_2^2\right] \leq \texttt{D}_Z,$$

with outer expectation over samples and training randomness. If $\tilde{s}(\widetilde{z}) = s_{\alpha^*}(\widetilde{z}) + (\widehat{\Sigma}_Z^{-1} + \epsilon I)\widetilde{z}$ for some $\epsilon > 0$, then $\widehat{\texttt{D}}_Z > 0$.

Theorem 4 provides a rigorous estimator $\widehat{\texttt{D}}_Z$ of multivariate mechanisms ($d \geq 1$). By training a score network via DSM on centered data, we derive $\widehat{\texttt{D}}_Z$ that provably upper-bounds half the Fisher information of Z and lower-bounds the true KL-divergence $\texttt{D}_Z$. The estimator is statistically efficient (unbiased in expectation), and regularization ensures strict positivity, enabling its use in the refinement of (worst-case) mutual information estimation.

Theorem 5 (Fourth-Cumulant Bound)**.** *Let X be any random variable, $\mathcal{M} : \mathcal{X} \to \mathbb{R}$ a scalar deterministic mechanism, and $B \sim \mathcal{N}(0, \sigma_B^2)$ independent Gaussian noise. Define $Z = \mathcal{M}(X) + B$ with mean $\mu_Z = \mathbb{E}[Z]$ and variance $\sigma_Z^2 = \mathrm{Var}(Z)$. Let $\texttt{D}_Z = D_{\mathrm{KL}}(P_Z \,\|\, \mathcal{N}(\mu_Z, \sigma_Z^2))$.*

Given i.i.d. samples $\{z_i\}_{i=1}^N$:

1) *Center data: $\widetilde{z}_i = z_i - \widehat{\mu}_Z$ where $\widehat{\mu}_Z = \frac{1}{N}\sum_{i=1}^N z_i$*
2) *Compute sample statistics: $\widehat{\sigma}_Z^2 = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_i^2$ and $\widehat{\kappa}_4 = \frac{1}{N}\sum_{i=1}^N \widetilde{z}_i^4 - 3(\widehat{\sigma}_Z^2)^2$.*

For any $c > 0$, define the estimator:

$$\widehat{\texttt{D}}_Z = \frac{\left(\max\left(\widehat{\kappa}_4, \frac{c}{N}\right)\right)^2}{48(\widehat{\sigma}_Z^2)^2}.$$

Then, (i) $\widehat{\texttt{D}}_Z > 0$ almost surely; (ii) Asymptotically: $\widehat{\texttt{D}}_Z \xrightarrow{a.s.} \frac{\kappa_4^2}{48\sigma_Z^4} \leq \texttt{D}_Z$ as $N \to \infty$; (iii) $\kappa_4 = \mathbb{E}[(Z - \mu_Y)^4] - 3\sigma_Z^4$ is the true excess kurtosis.

Theorem 5 provides a computationally efficient estimator $\widehat{\texttt{D}}_Z$ for scalar mechanism ($d = 1$) using excess hurtosis. By centering data and applying a max-operator, we guarantee $\widehat{\texttt{D}}_Z > 0$ almost surely while maintaining asymptotic convergence to the leading Edgeworth term $\frac{\kappa_4^2}{48\sigma_Z^4}$, which lower-bounds $\widehat{\texttt{D}}_Z$. The estimator requires only simple moments, avoids density estimation, and delivers consistent improvement over the conservative bound $\texttt{LogDet}(\mathcal{M}(X), B)$.

Corollary 2 directly follows Theorem 4 and 5.

Corollary 2. *Let $\mathcal{M} : \mathcal{X} \mapsto \mathbb{R}^d$ be an arbitrary deterministic mechanism and $B \sim \mathcal{N}(0, \Sigma_B)$ such that*

$\mathcal{M}(X) + B$ *has* $\texttt{LogDet}(\mathcal{M}(X), B) = \beta$*. Then, Under the assumptions of Theorems 4 and 5, the perturbed mechanism* $Z = \mathcal{M}(X) + B$ *is PAC private with*

$$\texttt{MI}(X; Z) \leq \beta - \widehat{\texttt{D}}_Z < \beta,$$

where $\widehat{\texttt{D}}_Z > 0$ *is obtained by Theorems 4 (for* $d \geq 1$*) or 5 (for* $d = 1$*).*

Corollary 3 shows that, once we account for the non-Gaussianity of the perturbed output through the positive correction term $\widehat{\texttt{D}}_Z$, the true mutual information always lies strictly below the nominal Gaussian bound. Concretely, if Z = M(X) + B is calibrated so that $\texttt{LogDet}\big(\mathcal{M}(X), B\big) = \beta$, then using either the Stein–discrepancy estimator (for $d \geq 1$) or the fourth-cumulant bound (for $d = 1$) yields a correction $\widehat{\texttt{D}}Z > 0$. As a result, $\texttt{MI}\big(X; Z\big) \leq \beta - \widehat{\texttt{D}}_Z < \beta$. In practice, this means that whenever the output distribution deviates from Gaussian, the term $\widehat{\texttt{D}}_Z$ exactly captures the "gap" between the surrogate bound and the true leakage. By subtracting $\widehat{\texttt{D}}_Z$ from β, one obtains a strictly tighter privacy guarantee, ensuring that no privacy budget is wasted on directions where Gaussianity-based estimates are overly conservative.

6. Stackelberg Automatic Residual-PAC Privatization

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

$$\begin{aligned} &\inf_Q \mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B; \mathcal{M})], \\ &\text{s.t. } \mathcal{H}(X|\mathcal{M}(X) + B) \geq \hat{\beta}, B \sim Q. \end{aligned} \tag{8}$$

By the definition of mutual information, any solution Q^* to problem (8) also solves (6) with PAC privacy budget $\beta = \mathcal{H}(X) - \hat{\beta}$. Given that mutual information and conditional entropy are related by $\texttt{MI}(X; \mathcal{M}(X) + B) = \mathcal{H}(X) - \mathcal{H}(X|\mathcal{M}(X) + B)$ where $\mathcal{H}(X)$ is fixed, solving the optimal perturbation problem (8) with conditional entropy constraints presents the same computational challenges as the mutual information formulation.

To address this limitation, we present novel automatic privatization approaches for Residual-PAC privacy, which we term *Stackelberg Automatic Residual-PAC Privatization (SR-PAC)*. Our approach is based on a Stackelberg game-theoretic characterization of the conditional-entropy-constrained optimization (8). We show that SR-PAC satisfies the Residual-PAC privacy guarantee and prove that it achieves optimal perturbation without a waste of the privacy budget. Consequently, when $\mathbb{E}_{Q,\mathcal{M},\mathcal{D}}[\mathcal{K}(B; \mathcal{M})] = \mathbb{E}_Q\big[\|B\|_2^2\big]$, algorithms implementing SR-PAC can achieve superior utility performance compared to the automatic PAC privatization algorithms described in Section 2.1, given the same mutual-information privacy budget.

6.1. A Stackelberg Game Model

Algorithm 3 Monte Carlo SR-PAC

Require: Privacy budget $\hat{\beta}$, parametrized 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 $\lambda, \phi \sim$ 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^*)

Our SR-PAC algorithm recasts the optimal perturbation problem (8) as a two-level Stackelberg game between a *Leader* (choosing the *perturbation rule* Q) and *Follower* (choosing the *decoder* attempting to infer X from Y). Let Γ denote a rich family of noise distributions. Let $\Pi = \{\pi : \pi(\cdot|y) \in \Delta(\mathcal{X}) \in \Delta(\mathcal{X}), y \in \mathcal{Y}\}$ 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 $\pi \in \Pi$ to minimize the expected log 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, $\pi^*(Q) \in \arg\inf_{\pi \in \Pi} W(Q, \pi)$.

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

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

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

$$\begin{cases} Q^* \in \arg\inf_{Q \in \Gamma} \mathbb{E}[\mathcal{K}(B; \mathcal{M})], \text{ s.t. } W\big(Q, \pi^*(Q)\big) \geq \hat{\beta}, \\ \pi^*(Q) \in \arg\inf_{\pi \in \Pi} W(Q, \pi). \end{cases} \tag{9}$$

When we consider output perturbation and the utility loss $\mathcal{K}$ is chosen such that $Q \mapsto \mathbb{E}_{X \sim P_X, B \sim Q}\big[\mathcal{K}(B; \mathcal{M})\big]$ is convex in Q, the problem (9) 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, 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 $Q \mapsto \mathbb{E}_{X\sim P_X, B\sim Q}\big[\mathcal{K}(B;\mathcal{M})\big]$ 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.

As stated in Proposition 4, a Stackelberg equilibrium perturbation rule solves (8).

Proposition 4. *Let (Q^*,π^*) be a Stackelberg equilibrium satisfying (9) for any $\hat{\beta}$. Then, Q^* solves (8) with privacy budget $\hat{\beta}$. In addition, in any Stackelberg equilibrium (Q^*,π^*), $\pi^* = \pi^*(Q^*)$ is unique.*

Algorithm 3 provides a Monte-Carlo-based approach to solve the Stackelberg equilibrium (9). By Monte Carlo sampling, this algorithm periodically trains the decoder to minimize reconstruction loss on perturbed data, enabling it to adapt to the current noise distribution. The perturbation rule is then optimized by balancing utility loss minimization against privacy constraints, using a penalty term that ensures the privacy cost remains close to the target budget.

6.2. 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. Our SR-PAC also supports anisotropic noise generation 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 $\langle \mathcal{M}(X), w\rangle$ 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 5. *Under Assumption 1, any Stackelberg-optimal 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 5 demonstrates that SR-PAC strategically allocates noise exclusively to the most privacy-sensitive directions. By adjusting the covariance matrix so that high-leakage dimensions receive proportionally more noise while low-leakage dimensions receive less, the method ensures optimal noise utilization where each unit of perturbation yields maximum privacy protection. This targeted approach enables SR-PAC to achieve the desired privacy level with minimal total perturbation, thereby preserving task-relevant information with significantly reduced distortion.

6.3. Directional-Selectivity of SR-PAC

Let $Z \in \mathbb{R}^d$ be an *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 $S_{\texttt{lab}} := \text{span}\{e_\ell - e_j : j \neq \ell\}$. 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 6. *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* $\sup_{v\in S_{\mathtt{lab}},\|v\|=1} g(v) < \inf_{w\perp S_{\mathtt{lab}},\|w\|=1} g(w)$. *Let* $\beta_{\mathtt{lab}} \equiv \frac{1}{2}\int_{w\perp S_{\mathtt{lab}}} g(w)d\sigma_w^2$ *be the maximal MI reduction achievable with noise* orthogonal *to* $S_{\mathtt{lab}}$. *Then, for every* $\beta \le \beta_{\mathtt{lab}}$, *we have*

$$\Pi_{\mathtt{lab}}B^\star = 0 \;\; a.s., \; \arg\max_i (Z_i + B_i^\star) = \hat{y} \;\; a.s.$$

In Proposition 6, part (i) shows that SR-PAC always uses strictly less noise power 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 $\beta_{\mathtt{lab}}$ is reached. In practice, this means SR-PAC perturbs only "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.

6.4. Sensitivity to β

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

Let

$$V_{\mathrm{SR}}(\beta) \equiv \min_{Q:\mathtt{MI}(X;\mathcal{M}(X)+B)\le\beta} \mathbb{E}_Q\left[\|B\|_2^2\right]$$

be the optimal noise-power curve attained by SR-PAC, and let $\mathtt{MI}_{\mathrm{SR}}(\beta)$ as the corresponding true mutual information attained by SR-PAC. Let $V_{\mathrm{PAC}}(\beta) \equiv \mathrm{tr}(\Sigma_{B_{\mathrm{PAC}}}(\beta))$, where $Q(\beta) = \mathcal{N}(0, \Sigma_{B_{\mathrm{PAC}}}(\beta))$ solves $\mathtt{LogDet}(\mathcal{M}(X), B_{\mathrm{PAC}}) = \beta$. In addition, let $\mathtt{MI}_{\mathrm{PAC}}(\beta) \equiv \beta - \mathtt{Gap}_{\mathtt{d}}(Q(\beta))$, where $\mathtt{Gap}_{\mathtt{d}}(Q) = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B}\|\widetilde{Q}_{\mathcal{M}})$ with $B \sim Q$, and $\widetilde{Q}_{\mathcal{M}}$ given by (4). Define $\mathtt{Priv}_\beta^{\mathrm{SR}} \equiv \frac{d}{d\beta}\mathtt{MI}_{\mathrm{SR}}(\beta)$, $\mathtt{Priv}_\beta^{\mathrm{PAC}} \equiv \frac{d}{d\beta}\mathtt{MI}_{\mathrm{PAC}}(\beta)$, $\mathtt{Util}_\beta^{\mathrm{SR}} \equiv \frac{d}{d\beta}(-V_{\mathrm{SR}}(\beta))$, and $\mathtt{Util}_\beta^{\mathrm{PAC}} \equiv \frac{d}{d\beta}(-V_{\mathrm{PAC}}(\beta))$.

Theorem 6. *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) $\mathtt{Priv}_\beta^{\mathrm{PAC}} \le \mathtt{Priv}_\beta^{\mathrm{SR}} = 1$, *with strict inequality for non-Gaussian* $\mathcal{M}(X)$.

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

Theorem 6 proves that SR-PAC with arbitrary noise distributions achieves: (i) *Exact leakage-budget alignment* ($\mathtt{Priv}_\beta^{\mathrm{SR}} = 1$), (ii) *Stricter utility decay* for Auto-PAC ($\mathtt{Util}_\beta^{\mathrm{SR}} \ge \mathtt{Util}_\beta^{\mathrm{PAC}}$). This holds unconditionally for non-Gaussian $\mathcal{M}(X)$ under privacy tightening (i.e., β decreasing).

Corollary 3. *In addition to the setting of Theorem 6, assume* $\varepsilon_{\mathrm{cal}}(\beta) \in [0, \mathtt{Gap}_{\mathtt{d}}\widehat{Q}(\beta))$, $\eta_{\mathrm{opt}}(\beta) \in [0, V_{\mathrm{PAC}}(\beta) - V_{\mathrm{SR}}(\beta))$. *Then, (i)* $|\mathtt{Priv}_\beta^{\mathrm{SR}} - 1| \le |\varepsilon'_{\mathrm{cal}}(\beta)|$; *(ii)* $\mathtt{Util}_\beta^{\mathrm{SR}} \ge \mathtt{Util}_\beta^{\mathrm{PAC}}$, *with equality only for Gaussian* $\mathcal{M}(X)$.

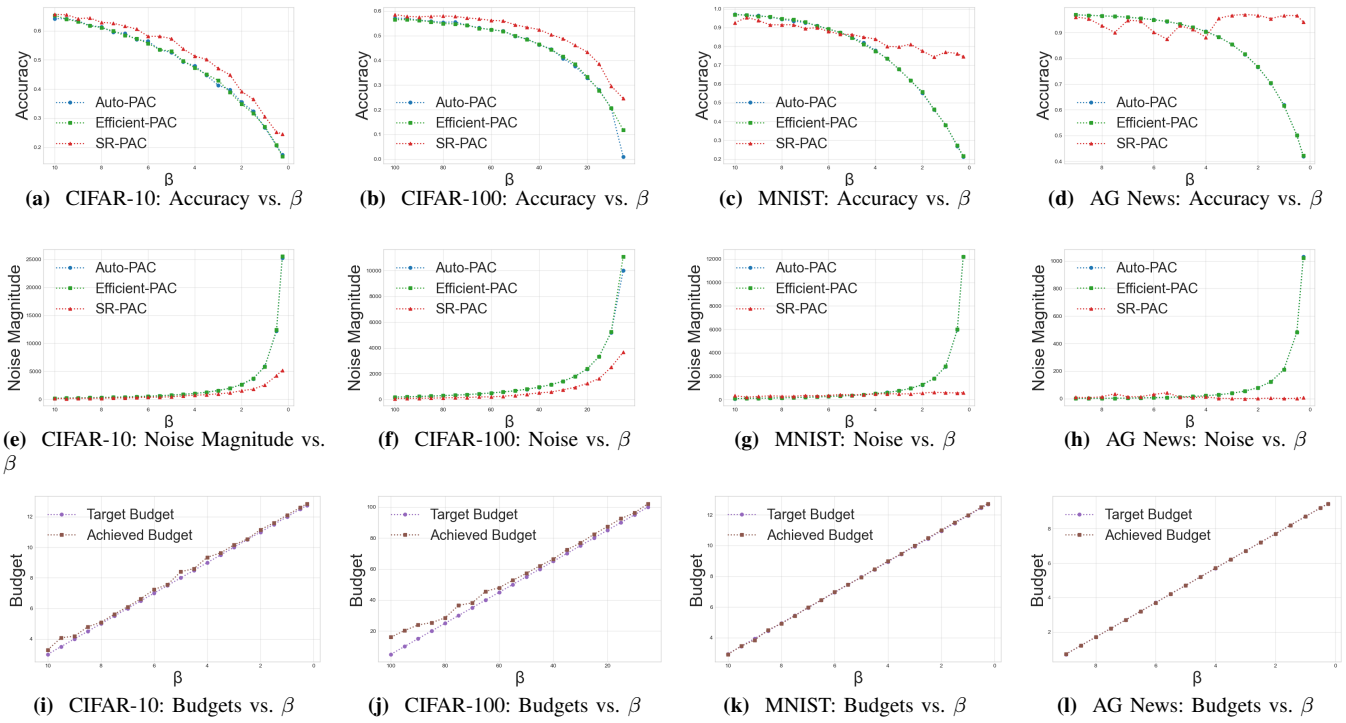
6.5. Composition

In this section, we study the composition properties of Residual-PAC Privacy and SR-PAC, focusing on the conditional entropy formulation. Strong composition properties are essential for privacy definitions like differential privacy, enabling systems to quantify aggregated privacy risk across sequential, adaptive, and concurrent operations on related datasets. This allows modular design where individual components maintain local privacy-utility trade-offs while keeping global privacy risk quantifiable.

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. 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 $\sum_{i=1}^k \beta_i$.

Residual-PAC Privacy also enjoys additive composition. Suppose each mechanism $\mathcal{M}_i$ is Residual-PAC private with conditional entropy lower bound $\hat{\beta}_i$. By definition of mutual information, 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 [1], 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 Residual-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 [1], and utility degradation compounds when each mechanism $\mathcal{M}_i$ uses conservative privacy budgets β_i. To address this limitation, we use an optimization-based approach within the SR-PAC framework to compute tighter conditional entropy bounds. Consider k mechanisms $\mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_k$ privatized by Q_i to satisfy Residual-PAC privacy with bounds $\hat{\beta}_i$. The Leader designs these



(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 2) on CIFAR-10, CIFAR-100, MNIST, and AG News as β varies. ach 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 power $\mathbb{E}[\|B\|_2^2]$ used by each method, and (bottom) the "target versus achieved" privacy budget for our SR-PAC.

privatizations $\{Q_i\}_{i=1}^k$, while the Follower finds the optimal decoder for the joint composition $\overrightarrow{\mathcal{M}}(X,\vec{\theta}) = (\mathcal{M}_1(X),\ldots,\mathcal{M}_k(X))$:

$$\inf_{\pi\in\Pi} W(\pi;\overrightarrow{\mathcal{M}}) \equiv \mathbb{E}_{X\sim\mathcal{D}}\left[-\log\pi(X|\overrightarrow{\mathcal{M}}(X),\vec{\theta})\right].$$

7. Numerical Experiments

In this experiments, we use the followng four datasets: CIFAR-10 [40], CIFAR-100 [40], MNIST dataset [41], and AG News dataset [42].

CIFAR-10 and Base Mechanism. The CIFAR-10 dataset comprises 50,000 training and 10,000 testing color images (each 32×32 pixels with three channels) divided evenly into ten classes (5,000 training and 1,000 testing images per class). Each image is converted to a $3 \times 32 \times 32$ tensor and normalized per channel to mean 0.5 and standard deviation 0.5. As the unperturbed mechanism, we train a convolutional neural network that consists of two convolutional blocks—each block is Conv→ReLU→MaxPool (kernel 2×2) with 32 filters in the first block and 64 filters in the second—followed by flattening into a 128-unit fully connected layer (with ReLU) and a final linear layer producing 10 logits. This network is trained by minimizing the cross-entropy loss over the CIFAR-10 classes. At inference, it maps each normalized image to a 10-dimensional logit vector, and the predicted label is given by the highest logit. The unperturbed mechanism achieves 0.7181 accuracy.

CIFAR-100 and Base Mechanism. CIFAR-100 contains 50,000 training and 10,000 testing color images (each $32 \times 32 \times 3$), equally divided among 100 fine-grained classes (500 training and 100 testing images per class). Each image is converted to a $3 \times 32 \times 32$ tensor and normalized per channel to mean 0.5 and standard deviation 0.5 before being fed into the network. As the unperturbed mechanism, we use a deeper convolutional neural network with three convolutional "blocks." Each block consists of two 3×3 convolutions (with BatchNorm and ReLU after each), followed by a 2×2 max-pool, which sequentially maps inputs from $32 \times 32 \to 16 \times 16 \to 8 \times 8 \to 4 \times 4$, with channel widths increasing from $3 \to 64 \to 128 \to 256$. After flattening the resulting $256 \times 4 \times 4$ feature map into a 4096-dimensional vector, a three-layer MLP head (4096→512→256→100) with ReLU activations and 0.5 dropout between the first two fully connected layers produces a 100-dimensional logit vector. During training, this network minimizes cross-entropy loss over the CIFAR-100 classes; at inference, each normalized image is mapped to its 100-dimensional logits, and the predicted label is given by the $\arg\max$ of those logits. The unperturbed mechanism achieves 0.5913 accuracy.

MNIST dataset and Base Mechanism. The MNIST dataset

comprises 60,000 training and 10,000 test grayscale images of handwritten digits (0–9). Each image is 28×28 pixels and is loaded as a $1 \times 28 \times 28$ tensor, then normalized to mean 0.1307 and standard deviation 0.3081 per channel before being fed into the network. As the unperturbed mechanism, we train a simple CNN consisting of two convolutional blocks—each block is Conv2d→BatchNorm→ReLU→MaxPool (2×2), with channel widths $1 \to 32 \to 64$—which produces a $64 \times 7 \times 7$ feature map. This feature map is flattened and passed through a two-layer fully connected head (128 units with ReLU+Dropout, then 10 output logits). At inference, each normalized 28×28 image is mapped to a 10-dimensional logit vector, and the predicted label is given by the index of the largest logit. The unperturbed mechanism achieves 0.9913 accuracy.

AG News dataset and Base Classifier. AG News comprises 120,000 training and 7,600 test articles equally divided among 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 into one text string, then lowercased and split on whitespace (truncated or padded to 64 tokens). We build a 30,000-word vocabulary from the training split and map each token to its index (with out-of-vocabulary tokens as 0). Those indices feed into an `nn.EmbeddingBag` layer (embedding size 300, mean-pooling mode) to produce a fixed-length 300-dimensional document vector. That vector is passed through a two-layer MLP head (300→256 with ReLU and 0.3 dropout, then 256→4), yielding a 4-dimensional logit vector, and at inference the predicted label is the index of the largest logit. The unperturbed mechanism achieves 0.9729 accuracy.

We evaluate SR-PAC against two baselines, Auto-PAC (Algorithm 1) and Efficient-PAC (Algorithm 2), on the above four datasets. For each dataset and its pretrained base mechanism $\mathcal{M}$, we plot (1) the test accuracy of the perturbed model as a function of β, (2) the average noise power $\mathbb{E}[\|B\|_2^2]$ required to achieve each β, and (3) SR-PAC's ability to hit the target budget $\widehat{\mathcal{H}}(X) - \beta$ (where $\widehat{H}(X)$ is our plug-in entropy estimate using the non-parametric k-nearest neighbour approach). Figure 1 summarizes these comparisons.

Recall that β is the user's desired upper bound on the mutual information $\mathtt{MI}(X; M(X) + B)$. By construction, SR-PAC enforces the equivalent conditional-entropy constraint $\mathcal{H}\big(X \mid M(X) + B\big) \geq \widehat{\beta} = \mathcal{H}(X) - \beta$, but a direct per-$\beta$ comparison of accuracy would require knowing the true data entropy $\mathcal{H}(X)$, which is unavailable in practice. Fortunately, for any fixed mechanism $\mathcal{M}$ and dataset there is a single "intrinsic" mutual information $\mathtt{MI}_o = \mathtt{MI}(X; M(X) + B)$, and no independent noise B can push $\mathtt{MI}(X; M(X) + B)$ above $\mathtt{MI}_o$. Equivalently, the only feasible budgets satisfy $0 < \beta \leq \mathtt{MI}_o$. At the left endpoint $\beta = \mathtt{MI}_o$, the best solution is $B = 0$, so all three methods coincide at the (near) noiseless accuracy. Although $\mathtt{MI}_o$ is unknown (and thus $\mathcal{H}(X)$ cannot be recovered), each algorithm still operates over the same interval $(0, \mathtt{MI}_o]$. In particular, every plotted β is a valid privacy target for Auto-PAC, Efficient-PAC, and SR-PAC. Therefore, even without knowing $\mathtt{MI}_o$ (or $\mathcal{H}(X)$), we can fairly compare all three curves—accuracy vs. β and noise magnitude vs. β—across the entire feasible range.

Accuracy vs. β (a-d of Figure 1): As β decreases (moving left along the horizontal axis), privacy increases. Thus, all three methods incur a drop in test accuracy. When β is relatively large (near $\mathtt{MI}o$), all three algorithms yield almost the same accuracy close to the noiseless accuracy. As β decreases, the SR-PAC curve always stays about the curves of Auto-PAC and Efficient-PAC—most dramatically on CIFAR-10 and CIFAR-100.

Noise Magnitude vs. (β (e-h) of Figure 1): As β decreases, each algorithm must add more noise, so all three curves rise. In every dataset, SR-PAC uses the (approximately) smallest $\mathbb{E}[\|B\|_2^2]$ at each β, whose curves never exceed the blue and green curves. Auto-PAC and Efficient-PAC both overshoot. That is, they add strictly more noise than SR-PAC.

The empirical ordering in both the accuracy and the noise magnitude results exactly matches Theorem 6. Note that the results in Theorem 6 hold for every non-Gaussian base mechanism. Figure 1 (c–d, g–h) shows that SR-PAC's performance on MNIST and AG News reflects the properties of Proposition 6. In particular, SR-PAC maintains nearly the same accuracy as the unperturbed mechanism over a wide range of budgets β, because for $\beta \leq \beta_{\mathtt{lab}}$ it injects all noise in directions orthogonal to the label subspace and thus never flips the predicted class. At the same time, its total noise magnitude remains an order of magnitude smaller than Auto-PAC or Efficient-PAC, since those conservatively calibrated schemes overestimate the required variance when the logits are highly non-Gaussian (MNIST and AG News both exhibit heavy-tailed logit distributions).

Budgets vs. β (i-l of Figure 1): These results show the SR-PAC's target vs achieved privacy budgets to verify that our algorithm indeed implement the specified desired privacy level. Each panel shows the target mutual-information bound β on the horizontal axis and SR-PAC's empirically measured one on the vertical axis. In every dataset, the red points lie nearly on the $y = x$ line. This confirms that SR-PAC solves its Followers' problem with high accuracy, so that the desired privacy budget is well guaranteed with negligible error, proving a reliable, data-driven guarantee that the privacy constraint is satisfied.

8. Conclusion

In this work, we introduced *Residual-PAC Privacy*, an enhanced framework that quantifies privacy guarantees beyond Gaussian assumptions and overcomes the inherent conservativeness of prior PAC-Privacy methods. By casting the privacy–utility trade-off as a convex Stackelberg optimization problem, our Stackelberg Residual-PAC (SR-PAC) approach fully leverages the available privacy budget and automatically calibrates anisotropic noise distributions tailored to the underlying data and mechanism. Extensive

numerical experiments on CIFAR-10, CIFAR-100, MNIST, and AG News confirm that SR-PAC consistently attains tighter privacy guarantees and higher utility compared to existing approaches. Consequently, Residual-PAC Privacy combines rigorous theory with practical effectiveness, offering a robust foundation for scalable, precise privacy assurance in complex, data-driven applications.

References

[1] H. Xiao and S. Devadas, "Pac privacy: Automatic privacy measurement and control of data processing," in *Annual International Cryptology Conference*. Springer, 2023, pp. 611–644.

[2] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*, 2006, pp. 1–12.

[3] A. D. P. Team *et al.*, "Learning with privacy at scale," *Apple Mach. Learn. J*, vol. 1, no. 8, pp. 1–25, 2017.

[4] J. M. Abowd, "The us census bureau adopts differential privacy," in *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, 2018, pp. 2867–2867.

[5] I. Issa, A. B. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, 2019.

[6] S. Saeidian, G. Cervia, T. J. Oechtering, and M. Skoglund, "Pointwise maximal leakage," *IEEE Transactions on Information Theory*, vol. 69, no. 12, pp. 8054–8080, 2023.

[7] P. Cuff and L. Yu, "Differential privacy as a mutual information constraint," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 43–54.

[8] F. Farokhi and H. Sandberg, "Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries," *IEEE Transactions on Smart Grid*, vol. 9, no. 5, pp. 4726–4734, 2017.

[9] A. Hannun, C. Guo, and L. van der Maaten, "Measuring data leakage in machine-learning models with fisher information," in *Uncertainty in Artificial Intelligence*. PMLR, 2021, pp. 760–770.

[10] C. Guo, B. Karrer, K. Chaudhuri, and L. van der Maaten, "Bounding training data reconstruction in private (deep) learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 8056–8071.

[11] X. Xiao and Y. Tao, "Output perturbation with query relaxation," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 857–869, 2008.

[12] J. Murtagh and S. Vadhan, "The complexity of computing the optimal composition of differential privacy," in *Theory of Cryptography Conference*. Springer, 2015, pp. 157–175.

[13] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP)*. IEEE, 2017, pp. 3–18.

[14] L. G. Valiant, "A theory of the learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

[15] M. Sridhar, H. Xiao, and S. Devadas, "Pac-private algorithms," *Cryptology ePrint Archive*, 2024.

[16] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.

[17] M. Bun and T. Steinke, "Concentrated differential privacy: Simplifications, extensions, and lower bounds," in *Theory of cryptography conference*. Springer, 2016, pp. 635–658.

[18] C. Dwork and G. N. Rothblum, "Concentrated differential privacy," *arXiv preprint arXiv:1603.01887*, 2016.

[19] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated cdp," in *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, 2018, pp. 74–86.

[20] I. Mironov, "Rényi differential privacy," in *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 2017, pp. 263–275.

[21] K. Chatzikokolakis, T. Chothia, and A. Guha, "Statistical measurement of information leakage," in *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*. Springer, 2010, pp. 390–404.

[22] G. Lebanon, M. Scannapieco, M. Fouad, and E. Bertino, "Beyond k-anonymity: A decision theoretic framework for assessing privacy risk," *Transactions on Data Privacy*, 2009.

[23] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[24] M. Lopuhaä-Zwakenberg and J. Goseling, "Mechanisms for robust local differential privacy," *Entropy*, vol. 26, no. 3, p. 233, 2024.

[25] A. Ghosh, T. Roughgarden, and M. Sundararajan, "Universally utility-maximizing privacy mechanisms," in *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009, pp. 351–360.

[26] M. Gupte and M. Sundararajan, "Universally optimal privacy mechanisms for minimax agents," in *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2010, pp. 135–146.

[27] Q. Geng, W. Ding, R. Guo, and S. Kumar, "Tight analysis of privacy and utility tradeoff in approximate differential privacy," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2020, pp. 89–99.

[28] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*. IEEE, 2012, pp. 1401–1408.

[29] W. Alghamdi, S. Asoodeh, F. P. Calmon, O. Kosut, L. Sankar, and F. Wei, "Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 1838–1843.

[30] J. Goseling and M. Lopuhaä-Zwakenberg, "Robust optimization for local differential privacy," in *2022 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2022, pp. 1629–1634.

[31] C. Huang, P. Kairouz, X. Chen, L. Sankar, and R. Rajagopal, "Generative adversarial privacy," *arXiv preprint arXiv:1807.05306*, 2018.

[32] X. Chen, P. Kairouz, and R. Rajagopal, "Understanding compressive adversarial privacy," in *2018 IEEE Conference on Decision and Control (CDC)*. IEEE, 2018, pp. 6824–6831.

[33] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, 2018, pp. 634–646.

[34] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International conference on learning representations*, 2018.

[35] A. Selvi, H. Liu, and W. Wiesemann, "Differential privacy via distributionally robust optimization," *Operations Research*, 2025.

[36] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer, "Membership inference attacks from first principles," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1897–1914.

[37] B. Balle, G. Cherubin, and J. Hayes, "Reconstructing training data with informed adversaries," in *2022 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2022, pp. 1138–1156.

[38] P. Vincent, "A connection between score matching and denoising autoencoders," *Neural computation*, vol. 23, no. 7, pp. 1661–1674, 2011.

[39] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[40] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.

[41] Y. LeCun, C. Cortes, and C. Burges, "Mnist handwritten digit database," *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, vol. 2, 2010.

[42] X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

[43] A. J. Stam, "Some inequalities satisfied by the quantities of information of fisher and shannon," *Information and Control*, vol. 2, no. 2, pp. 101–112, 1959.

[44] A. R. Barron, "Entropy and the central limit theorem," *The Annals of probability*, pp. 336–342, 1986.

[45] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector gaussian channels," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 141–154, 2005.

[46] S. Park, E. Serpedin, and K. Qaraqe, "On the equivalence between stein and de bruijn identities," *IEEE Transactions on Information Theory*, vol. 58, no. 12, pp. 7045–7067, 2012.

Appendix A. Proof of Proposition 2

Proof. Since $B \sim \mathcal{N}(0, \Sigma_B)$, we have $\mathbb{E}[\|B\|_2^2] = \mathrm{tr}\big(\mathbb{E}[B\,B^\top]\big) = \mathrm{tr}(\Sigma_B)$. Hence, minimizing $\mathbb{E}[\|B\|_2^2]$ over zero-mean Gaussian is equivalent to minimizing the trace $\mathrm{tr}(\Sigma_B)$ over $\Sigma_B \succeq 0$.

Recall that $Z = \mathcal{M}(X) + B$. Then, Z has mean $\mu_Z = \mu_{\mathcal{M}(X)}$ and covariance $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$, where $\Sigma_{\mathcal{M}(X)}$ denotes the covariance of $\mathcal{M}(X)$. In addition, recall that $\widetilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$ is the Gaussian distribution with the same first and second moments as Z. Then, by standard Gaussian-entropy formulas, we have

$$\begin{aligned}\texttt{MI}(X;Z) = H(Z) - H(Z|X) &= \frac{1}{2}\log\frac{\det(\Sigma_Z)}{\det(\Sigma_B)} \\ &= \frac{1}{2}\log\det(I + \Sigma_{\mathcal{M}(X)}\Sigma_B^{-1}).\end{aligned}$$

In particular, Algorithm 1 implements $\texttt{MI}(X;Z) \le \beta$.

Since both $\mathrm{tr}(\Sigma_B)$ and $\log\det(I + \Sigma_{\mathcal{M}(X)}\Sigma_B^{-1})$ are unitarily invariant, we may diagonalize $\Sigma_{\mathcal{M}(X)}$ as

$$\Sigma_{\mathcal{M}(X)} = U\mathrm{diag}(r_1, \ldots, r_d)U^T,\ r_i > 0,$$

where U is the orthogonal eigenvector matrix from the eigendecomposition of $\Sigma_{\mathcal{M}(X)}$. Writing $\Sigma_B = \hat{U}\mathrm{diag}(\ell_1, \ldots, \ell_d)\hat{U}^T$ with $\ell_i > 0$, the problem

$$\min_{\Sigma_B \succeq 0} \mathrm{tr}(\Sigma_B),\ \text{s.t.}\ \frac{1}{2}\log\det(1 + \Sigma_{\mathcal{M}(X)}\Sigma_B^{-1}) = \beta,$$

becomes

$$\min_{\ell_1,\ldots,\ell_d > 0} \sum_{i=1}^d \ell_i,\ \text{s.t.}\ \frac{1}{2}\sum_{i=1}^d \log(1 + \frac{r_i}{\ell_i}) = \beta.$$

Hence, each coordinate ℓ_i appears only in the term $\log(1 + \frac{r_i}{\ell_i})$.

Let $\lambda > 0$ as the Lagrange multiplier. The Lagrangian is

$$\begin{aligned}&\mathcal{L}(\ell_1, \ldots, \ell_d, \lambda) \\ &= \sum_{i=1}^d \ell_i + \lambda\left(\frac{1}{2}\sum_{i=1}^d \log(1 + \frac{r_i}{\ell_i})\texttt{MI}(X;\widetilde{Z}) - \beta\right).\end{aligned}$$

Setting $\frac{\partial\mathcal{L}}{\partial\ell_i} = 0$ gives $1 = \lambda\frac{r_i}{2\ell_i(\ell_i + r_i)} \Rightarrow 2\ell_i(\ell_i, r_i) = \lambda r_i$. Equivalently, $\ell_i^2 + r_i\ell_i - \lambda\frac{r_i}{2} = 0$, which gives a unique $\ell_i(\lambda) = \frac{-r_i + \sqrt{r_i^2 + 2\lambda r_i}}{2} > 0$.

Let $F(\lambda) = \frac{1}{2}\sum_{i=1}^d \log(1 + \frac{r_i}{\ell_i(\lambda)})$. We can check that as $\lambda \to 0^+$, each $\ell_i(\lambda) \to 0^+$, so $F(\lambda) \to +\infty$. As $\lambda \to +\infty$, each $\ell_i(\lambda) \to +\infty$, leading to $F(\lambda) \to +\infty$. In addition, $\frac{dF(\lambda)}{d\lambda} < 0$ throughout, so F is strictly decreasing from $+\infty$ down to 0. Therefore, there is a unique $\lambda^* > 0$ such that $F(\lambda^*) = \beta$. At this λ^*, each $\ell_i^* = \ell_i^*(\lambda^*)$ is unique. Thus, $\Sigma_B^* = \hat{U}\mathrm{diag}(\ell_1^*, \ldots, \ell_d^*)\hat{U}^T$ is unique minimizer of $\mathrm{tr}(\Sigma_B)$. By construction, $\frac{1}{2}\log\det\left(1 + \Sigma_{\mathcal{M}(X)}(\Sigma_B^*)^{-1}\right) = \beta$. Therefore, it is also the unique minimizer of (5). □

Appendix B. Proof of Theorem 3

By Lemma 1 (which is shown and proved later), the function $g(\beta) = \texttt{Gap}_\texttt{d}(Q^*(\beta))$ is nondecreasing in β. Thus, for any $0 < \beta_1 < \beta_2$, we have $\texttt{Gap}_\texttt{d}(Q^*(\beta_2)) \ge \texttt{Gap}_\texttt{d}(Q^*(\beta_1))$, which yields $\texttt{G}(\beta_2, \beta_1) = \texttt{Gap}_\texttt{d}(Q^*(\beta_2)) - \texttt{Gap}_\texttt{d}(Q^*(\beta_1)) \ge 0$. Recall the relationship between true mutual information and the bound $\texttt{LogDet}(\mathcal{M}(X), B) = \beta$:

$$\texttt{I}_{\mathrm{true}}\big(Q^*(\beta)\big) = \beta\ -\ \texttt{Gap}_\texttt{d}\big(Q^*(\beta)\big).$$

Hence, for $0 < \beta_1 < \beta_2$,

$$\begin{aligned}&\texttt{I}_{\mathrm{true}}\big(Q^*(\beta_2)\big)\ -\ \texttt{I}_{\mathrm{true}}\big(Q^*(\beta_1)\big) \\ &= \big[\beta_2 - \texttt{Gap}_\texttt{d}(Q^*(\beta_2))\big] - \big[\beta_1 - \texttt{Gap}_\texttt{d}(Q^*(\beta_1))\big] \\ &= (\beta_2 - \beta_1)\ -\ \big[\texttt{Gap}_\texttt{d}(Q^*(\beta_2)) - \texttt{Gap}_\texttt{d}(Q^*(\beta_1))\big] \\ &= (\beta_2 - \beta_1)\ -\ \texttt{G}(\beta_2, \beta_1).\end{aligned}$$

The two bullet points now follow immediately: (i) If $\texttt{G}(\beta_2, \beta_1) \le \beta_2 - \beta_1$, then

$$\texttt{I}_{\mathrm{true}}(Q^*(\beta_2)) - \texttt{I}_{\mathrm{true}}(Q^*(\beta_1)) = (\beta_2 - \beta_1) - \texttt{G}(\beta_2, \beta_1) \ge 0,$$

i.e. $\texttt{I}_{\mathrm{true}}(Q^*(\beta_1)) \le \texttt{I}_{\mathrm{true}}(Q^*(\beta_2))$. (ii) If $\texttt{G}(\beta_2, \beta_1) > \beta_2 - \beta_1$, then

$$\texttt{I}_{\mathrm{true}}(Q^*(\beta_2)) - \texttt{I}_{\mathrm{true}}(Q^*(\beta_1)) = (\beta_2 - \beta_1) - \texttt{G}(\beta_2, \beta_1) < 0,$$

i.e. $\texttt{I}_{\mathrm{true}}(Q^*(\beta_1)) > \texttt{I}_{\mathrm{true}}(Q^*(\beta_2))$.

□

Lemma 1. *Fix a mechanism $\mathcal{M}$ and a data distribution $\mathcal{D}$. Let $Q^*(\beta)$ be the solution of (5). Then, $\texttt{Gap}_\texttt{d}(Q^*(\beta))$ is a nondecreasing function of β.*

Proof. Let $g(\beta) = \mathtt{Gap}_{\mathtt{d}}(Q^*(\beta))$ and $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$ with $\Sigma_B = \Sigma_B^*(\beta)$. By definition,

$$g(\beta) = H\big(\mathcal{N}(0, \Sigma_Z)\big) - H\big(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B)\big).$$

Differentiate with respect to β via the chain rule:

$$\frac{dg}{d\beta} = \left\langle \nabla_{\Sigma_B} \left[H(\mathcal{N}(0, \Sigma_Z)) - H(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B)) \right], \frac{d\Sigma_B}{d\beta} \right\rangle.$$

The gradient of Gaussian entropy is $\nabla_{\Sigma_B} H(\mathcal{N}(0, \Sigma_Z)) = \frac{1}{2}\Sigma_Z^{-1}$. By de Bruijn's identity [43],

$$\nabla_{\Sigma_B} H(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B)) = \frac{1}{2} J(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B)),$$

where $J(\cdot)$ is the Fisher information. The Cramér–Rao bound gives $J(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B)) \succeq \Sigma_Z^{-1}$. Thus,

$$\nabla_{\Sigma_B} g = \frac{1}{2}\left(\Sigma_Z^{-1} - J(P_{\mathcal{M}} * \mathcal{N}(0, \Sigma_B))\right) \preceq 0.$$

From Proposition 2, $\frac{d\Sigma_B}{d\beta} \preceq 0$ (strictly negative when Σ_B changes). Since both $\nabla_{\Sigma_B} g$ and $\frac{d\Sigma_B}{d\beta}$ are symmetric negative semidefinite,

$$\frac{dg}{d\beta} = \left\langle \nabla_{\Sigma_B} g, \frac{d\Sigma_B}{d\beta} \right\rangle = \operatorname{tr}\left((\nabla_{\Sigma_B} g) \left(\frac{d\Sigma_B}{d\beta} \right) \right) \geq 0,$$

as the trace of the product of two negative semidefinite matrices is nonnegative. Hence $g(\beta)$ is nondecreasing. □

Appendix C. Proof of Theorem 4

We first prove the inequality chain, and then prove the strict positivity $\widehat{\mathtt{D}}_Z > 0$.

C.1. Part 1: Inequality Chain

Define the Stein score

$$T(\tilde{Z}) = \nabla_{\tilde{z}} \log P_{\widetilde{Z}}(\tilde{z}) + \Sigma_{\widetilde{Z}}^{-1} \tilde{z}.$$

For any distribution P on $\mathbb{R}^d$ with density p, mean μ, and covariance Σ, and the moment-matched Gaussian $g = \mathcal{N}(\mu, \Sigma)$, the Gaussian log-Sobolev inequality gives

$$\mathtt{D}_{\mathrm{KL}}(P\|g) \geq \frac{1}{2}\mathbb{E}_P\left[\|T(Z_p)\|_2^2\right] \equiv \frac{1}{2}\mathcal{J}(P\|g), \tag{10}$$

where $Z_p \sim P$.

Let s_{α^*} be the DSM-optimal score estimator trained on the centered samples $\{\tilde{z}_i\}_{i=1}^N$. By [38] and [39], $s_{\alpha^*}(\tilde{z})$ converges to $\Delta_{\tilde{z}} \log P_{\widetilde{Z}}(\tilde{z})$ in $L^2(P_{\widetilde{Z}})$ as $N \to \infty$ and $\epsilon \to 0$.

Taking outer expectation over the i.i.d. draws of the sample $\{\tilde{z}_i\}$ and the DSM training randomness yields

$$\mathbb{E}[\widehat{\mathtt{D}}_Z] = \frac{1}{2}\mathbb{E}_{\widehat{Z}}\left[\|s_{\alpha^*}(\widehat{Z})\|_2^2\right] - \mathbb{E}_{\widehat{Z}}\left[\nabla \cdot s_{\alpha^*}(\widetilde{Z})\right].$$

For any square0integrable vector filds f, by the Stein' identity, we have

$$\mathbb{E}_{\widetilde{Z}}\left[\nabla \cdot f(\widetilde{Z})\right] = -\mathbb{E}_{\widetilde{Z}}\left[f(\widetilde{Z})^\top \Sigma_{\widetilde{Z}}^{-1} \widetilde{Z}\right].$$

When $f = s_{\alpha^*}$, we get

$$\mathbb{E}[\widehat{\mathtt{D}}_Z] = \frac{1}{2}\mathbb{E}_{\widehat{Z}}\left[\|s_{\alpha^*}(\widehat{Z})\|_2^2 + 2 s_{\alpha^*}^\top \Sigma_{\widetilde{Z}}^{-1} \widetilde{Z}\right].$$

Adding and subtracting $\frac{1}{2}\|\Sigma_{\widetilde{Z}}^{-1/2}\widetilde{Z}\|_2^2$ gives

$$\mathbb{E}\left[\widehat{\mathtt{D}}_Z\right] = \frac{1}{2}\mathbb{E}\left[\|s_{\alpha^*}(\widetilde{Z}) + \Sigma_{\widetilde{Z}}^{-1}\widetilde{Z}\|_2^2\right] - \frac{1}{2}\mathbb{E}\left[\|\Sigma_{\widetilde{Z}}^{-1/2}\widetilde{Z}\|_2^2\right].$$

Since $\mu_{\widetilde{Z}} = 0$ and $\Sigma_{\widetilde{Z}} = \Sigma_Z \succ 0$,

$$\frac{1}{2}\mathbb{E}\left[\|\Sigma_{\widetilde{Z}}^{-1/2}\widetilde{Z}\|_2^2\right] = \frac{d}{2}$$

is a constant that is independent of s_{α^*}. Thus, we have

$$\mathbb{E}\left[\widehat{\mathtt{D}}_Z\right] \leq \frac{1}{2}\mathbb{E}\left[\|s_{\alpha^*}(\widetilde{Z}) + \Sigma_{\widetilde{Z}}^{-1}\widetilde{Z}\|_2^2\right] = \frac{1}{2}\mathbb{E}\left[\|T_{\alpha^*}(\widetilde{Z})\|_2^2\right],$$

where $T_{\alpha^*}(\tilde{z}) \equiv s_{\alpha^*}(\tilde{z}) + \Sigma_{\widetilde{Z}}^{-1}\tilde{z}$.

Therefore, as $N \to \infty$ and $\epsilon \to 0$, we have

$$\lim_{N\to\infty, \epsilon\to 0} \mathbb{E}\left[\widehat{\mathtt{D}}_Z\right] \leq \frac{1}{2}\mathbb{E}\left[\|T(\widetilde{Z})\|_2^2\right] = \frac{1}{2}\mathcal{J}(P_{\widetilde{Z}}\|\mathcal{N}).$$

By (10), $\frac{1}{2}\mathcal{J}(P_{\widetilde{Z}}\|\mathcal{N}) \leq \mathtt{D}_Z$. Therefore, we obtain the inequality chain

$$\mathbb{E}\left[\widehat{\mathtt{D}}_Z\right] \leq \frac{1}{2}\mathbb{E}\left[\|\nabla_z \ln p_Z(Z)\|_2^2\right] \leq \mathtt{D}_Z.$$

C.2. Part 2: Strict Positivity

For $\epsilon > 0$, define

$$\tilde{s}(\tilde{z}) = s_{\alpha^*}(\widetilde{z}) + (\widehat{\Sigma}_Z^{-1} + \epsilon I)\widetilde{z}$$

Then, $T_{\tilde{s}}(\tilde{z}) = \tilde{s}(\tilde{z}) + \widehat{\Sigma}_Z^{-1}\tilde{z}$ differs from zero by at least the term $\epsilon\tilde{z}$. Unless in the degenerate case when $\widehat{Z}$ is almost surely zero,

$$\mathbb{E}[\|T_{\tilde{s}}(\widetilde{Z})\|_2^2] \geq \epsilon^2 \mathbb{E}[\|\widetilde{Z}\|_2^2] > 0.$$

Here, $\mathbb{E}[\|T_{\tilde{s}}(\widetilde{Z})\|_2^2] \geq d$, where the equality holds only for Gaussian $\widetilde{Z}$. Then, from

$$\widehat{\mathtt{D}}_Z = \frac{1}{2}\|T_{\tilde{s}}(\tilde{z})\|_2^2 - \frac{d}{2},$$

we guarantee $\widehat{\mathtt{D}}_Z > 0$ for non-Gaussian Z.

□

Appendix D. Proof of Theorem 5

For any real-valued random variable Z with mean μ_z and variance $\sigma_Z^2 < \infty$, write the excess kurtosis,

$$\kappa_4 = \mathbb{E}[(Z - \mu_Z)^2] - 3\sigma_Z^4.$$

By the fourth-moment bund [44], the KL divergence between P_Z and the moment-matched Gaussian satisfies

$$\mathtt{D}_{\mathrm{KL}}(P_Z\|\mathcal{N}(\mu_Z, \sigma_Z^2)) \geq \frac{\kappa^2}{48\sigma_Z^4}. \tag{11}$$

With i.i.d. samples $\{z_i\} = 1$, define the centered samples $\tilde{z}_i = z_i - \mu_z$ and the empirical moments

$$\tilde{\sigma}_Z = \frac{1}{N}\sum_{i=1}^{N} \tilde{z}_i^2,\ \tilde{\kappa}_4 = \frac{1}{N}\sum_{i=1}^{N} \tilde{z}_i^2 - 3(\tilde{\sigma}_Z^2)^2.$$

By the strong law of large numbers, we have

$$\tilde{\sigma}_Z \xrightarrow{\text{a.s.}} \sigma_Z^2,\ \tilde{\kappa}_4 \xrightarrow{\text{a.s.}} \kappa_4.$$

For a fixed $c > 0$, let $\widehat{\mathtt{D}}_Z = \frac{\left(\max\left(\tilde{\kappa}_4, \frac{c}{N}\right)\right)^2}{48(\tilde{\sigma}_Z^2)^2}$, where the numerator is the square of a non-negative quantity and the denominator is a positive empirical variance. Thus, $\widehat{\mathtt{D}}_Z > 0$ almost surely for every finite N.

By $\tilde{\sigma}_Z \xrightarrow{\text{a.s.}} \sigma_Z^2$, $\tilde{\kappa}_4 \xrightarrow{\text{a.s.}} \kappa_4$, and the continuous mapping theorem, we have $\widehat{\mathtt{D}}_Z \xrightarrow{\text{a.s.}} \frac{\kappa_4^2}{48\sigma_Z^4}$. From (11), $\widehat{\mathtt{D}}_Z \le \widehat{D}_Z$. Therefore, the estimator $0 < \widehat{\mathtt{D}}_Z \le \widehat{D}_Z$. □

Appendix E. Proof of Proposition 4

Fix any Q. The Follower's problem is to find $\pi^*(Q)$ solving $\inf_{\pi\in\Pi} W(Q,\pi)$. By definition

$$W(Q,\pi) = \mathbb{E}_{X\sim\mathcal{D},B\sim Q}\left[-\log \pi(X|\mathcal{M}(X)+B)\right]$$
$$-\int_{\mathcal{X},\mathcal{Y},\mathcal{R}^d} P_X(x)G_{\mathcal{M},Q}(y|x,b)\log\pi(x|y+b)dxdydb,$$

where $P_X(x)$ is the density function associated with data distribution $\mathcal{D}$, and $G_{\mathcal{M},Q}(y|x,b)$ is the conditional density function given $\mathcal{M}$ and Q.

Let $\eta_Q : \mathcal{Y} \mapsto \Delta(\mathcal{X})$ denote the posterior distirbution given P_X and $G_{\mathcal{M},Q}$. For any $\pi \in \Pi$, consider

$$W(Q,\pi) - W(Q,\eta_Q)$$
$$= \int_{\mathcal{X},\mathcal{Y},\mathcal{R}^d} P_X(x)G_{\mathcal{M},Q}(y|x,b)\log\eta_Q(x|y+b)dxdydb$$
$$-\int_{\mathcal{X},\mathcal{Y},\mathcal{R}^d} P_X(x)G_{\mathcal{M},Q}(y|x,b)\log\pi(x|y+b)dxdydb$$
$$= \int_{\mathcal{X},\mathcal{Y},\mathcal{R}^d} P_X(x)G_{\mathcal{M},Q}(y|x,b)\log\frac{\eta_Q(x|y+b)}{\pi(x|y+b)}dxdydb.$$

Let

$$\mathbf{P}_Q(y) \equiv \int_{\mathcal{X},\mathcal{R}^d} P_X(x)G_{\mathcal{M},Q}(y|x,b)dxdb.$$

By definition, we have

$$\eta_Q\mathbf{P}_Q(y) = \int_{\mathcal{X}} P_X(x)G_{\mathcal{M},Q}(y|x,b).$$

Thus, for all $Q \in \Gamma$,

$$W(Q,\pi) - W(Q,\eta_Q) = \mathtt{D}_{\mathrm{KL}}(\eta_Q\|\pi) \ge 0.$$

Then, $W(Q,\pi) \ge W(Q,\eta_Q)$, where the equality holds if and only if $\pi = \eta_Q$. That is, for any $Q \in \Gamma$, there is a unique $\pi(Q)$ as a solution of $\inf_{\pi\in\Pi} W(Q,\pi)$. In addition, when $\pi(Q) = \eta_Q$, $W(Q,\pi(Q))$ is the conditional entropy. □

Appendix F. Proof of Proposition 5

Based on (iii) of Assumption 1, consider $K(Q) = \mathbb{E}_{B\sim Q}\big[g(\|B\|)\big]$, where $g : \mathbb{R}_+ \to \mathbb{R}$ is strictly increasing and strictly convex.

Suppose, to reach a contradiction, that an optimal $Q^\star$ is isotropic with $\Sigma_{Q^\star} = \sigma^2 I_d$ and attains the constraint with equality: $\mathcal{H}(X \mid \mathcal{M}(X)+B) = \hat{\beta}$.

For small $\Delta_v > 0$ define the perturbed covariance

$$\Sigma'(\Delta_v) \equiv (\sigma^2 - \Delta_v)\,vv^\top + (\sigma^2 + \Delta_u)\,uu^\top + \sigma^2 P_{\{u,v\}^\perp},$$

with $\Delta_u \in (0, \Delta_v)$ to be chosen. Denote by $\mathtt{h}(\sigma_u^2, \sigma_v^2) \equiv \mathcal{H}(X|Y)$ the conditional entropy evaluated at those directional variances.

Because $\mathtt{h}$ is C^1 and strictly increasing in each argument, we have

$$\left.\frac{\partial \mathtt{h}}{\partial\sigma_u^2}\right|_{\sigma^2} > \left.\frac{\partial \mathtt{h}}{\partial\sigma_v^2}\right|_{\sigma^2} > 0.$$

Hence the map

$$\phi_{\Delta_v}(\Delta_u) \equiv h(\sigma^2 + \Delta_u, \sigma^2 - \Delta_v)$$

is continuous and strictly increasing near $\Delta_u = 0$, with

$$\phi_{\Delta_v}(0) = \hat{\beta} - \frac{\partial h}{\partial\sigma_v^2}\Delta_v + o(\Delta_v) < \hat{\beta}.$$

By the Intermediate Value Theorem, there exists a unique $\Delta_u \in (0, \Delta_v)$ such that $\phi_{\Delta_v}(\Delta_u) = \hat{\beta}$, i.e. the perturbed noise Q' *satisfies the privacy constraint exactly*.

Because g is strictly convex,

$$g(\sigma^2 + \Delta_u) - g(\sigma^2) < g'(\sigma^2)\,\Delta_u,$$
$$g(\sigma^2 - \Delta_v) - g(\sigma^2) > g'(\sigma^2)(-\Delta_v).$$

Therefore $\mathcal{K}(Q') - \mathcal{K}(Q^\star) < g'(\sigma^2)(\Delta_u - \Delta_v) < 0$. That is, Q' is feasible and cheaper than $Q^\star$, contradicting optimality. Hence no optimum can be isotropic, so every minimiser must have $\lambda_{\max}(\Sigma) > \lambda_{\min}(\Sigma)$.

□

Appendix G. Proof of Proposition 6

G.1. Part (i):

Since entropy is maximised by a Gaussian with fixed covariance, the entropy-power inequality give

$$\mathcal{H}(Z + B_{\mathtt{pac}}) < \mathcal{H}(Z_{\mathrm{G}} + B_{\mathtt{pac}}),$$

where Z_{G} is Gaussian with covariance Σ_Z. Thus, $\mathtt{MI}(Z; Z + B_{\mathtt{pac}}) < \mathtt{MI}(Z_{\mathrm{G}}; Z_{\mathrm{G}} + B_{\mathtt{pac}}) = \beta$. To raise the mutual information back up to β, we can strictly reduce every directional variance of $B_{\mathtt{pac}}$. The optimizer Q^* therefore expands strictly less power. That is, $\mathbb{E}_{Q^*}[\|B\|_2^2] < \mathbb{E}[\|B_{\mathtt{pac}}\|_2^2]$.

G.2. Part (ii):

Let $\sigma_w^2 \equiv \mathtt{Var}\langle B, w\rangle$. Form the Lagrangian

$$\mathcal{L}(Q,\lambda) = \mathbb{E}_Q[\|B\|_2^2] + \lambda\big(\mathtt{MI}(Z; Z+B) - \beta\big).$$

For the stationarity condition w.r.t. each σ_w^2 we need the gradient of mutual information. By [45], we have

$$\partial_{\sigma_w^2}\mathtt{MI}(Z; Z+B) = g(w).$$

Hence $\partial_{\sigma_w^2}\mathcal{L} = 1 + \lambda\, g(w)$. The KKT conditions therefore read

$$1+\lambda g(w) = 0 \quad \text{if } \sigma_w^2 > 0,\ 1+\lambda g(w) \geq 0 \quad \text{if } \sigma_w^2 = 0,$$

for a unique $\lambda < 0$. Under the assumption

$$\sup_{v\in S_{\mathtt{lab}}, \|v\|=1} g(v) < \inf_{w\perp S_{\mathtt{lab}}, \|w\|=1} g(w),$$

these equalities can hold only for as long as the required mutual information reduction does not exceed $\beta_{\mathtt{lab}}$. Therefore, $\sigma_v^2 = 0$ for every $v \in S_{\mathtt{lab}}$. With those label-directions undisturbed, each class margin $e_\ell - e_j$ retains its sign, whence $\arg\max_i(Z_i + B_i^\star) = \hat{y}$.

□

Appendix H. Proof of Theorem 6

H.1. Part (i)

Let $Z = \mathcal{M}(X) + B$. For SR-PAC, the perurbation rule Q_{SR} satisfies $\mathcal{H}(X|Z) = \mathcal{H}(X) - \beta$. By definition, we have $\mathtt{MI}_{\mathrm{SR}}(\beta) = \beta$. Thus, $\mathtt{Priv}_\beta^{\mathrm{SR}} = 1$.

For Auto-PAC, the noise $B_{\mathrm{PAC}} \sim \mathcal{N}(0, \Sigma_{B_{\mathrm{PAC}}}(\beta))$ satisfies $\frac{1}{2}\log\det\big(I_d + \Sigma_{\mathcal{M}(X)}\Sigma_{B_{\mathrm{PAC}}}^{-1}(\beta)\big) = \beta$. By Proposition 1, the true mutual information is

$$\mathtt{MI}_{\mathrm{PAC}}(\beta) = \beta - \mathtt{Gap}_d(\beta),$$

where $\mathtt{Gap}_{\mathtt{d}}(\beta) = \mathtt{D}_{\mathrm{KL}}(P_{\mathcal{M},B_{\mathrm{PAC}}}\|\widetilde{Q}_{\mathcal{M}}) \geq 0$. When $\mathcal{M}(X)$ is non-Gaussian, $\mathtt{Gap}_{\mathtt{d}}(\beta) > 0$ for all $\beta > 0$. By de Bruijin's idensity (e.g., [46]),

$$\frac{d}{d\beta}\mathtt{Gap}_{\mathtt{d}}(\beta) = \frac{1}{2}\mathcal{J}(P_{\mathcal{M}+B_{\mathrm{PAC}]}}(\beta)\|\widetilde{Q}_{\mathcal{M}}) > 0,$$

where $\mathcal{J}(\cdot\|\cdot)$ is the relative Fisher information. Thus, $\mathtt{Priv}_\beta^{\mathrm{PAC}} = \frac{d}{d\beta}\mathtt{MI}_{\mathrm{PAC}}(\beta) < 1 = \mathtt{MI}_{\mathrm{SR}}(\beta)$.

H.2. Part (ii)

It is well known that for a fixed prior, mutual information is convex in the channel law. When $Z = \mathcal{M}(X) + B$, the "channel law" in our setting of the deterministic mechanism is determined by the perturbation rule Q. Thus, the mapping $Q \mapsto \mathtt{MI}(Q) \equiv \mathtt{MI}(X; \mathcal{M}(X) + B$ is convex. The objective $\mathcal{K}(Q) = \mathbb{E}_Q[\|B\|_2^2]$ is linear (hence convex) in Q. In addition, the constraint set $\{Q : \mathtt{MI}(Q) \leq \beta\}$ is convex. Then, Slater's condition holds because:

(i) when $\Sigma_B \to \infty$, $\mathtt{MI}(Q) \to 0 < \beta$;
(ii) $V(\beta)$ is finite for all $\beta > 0$ since $\mathbb{E}[\|\mathcal{M}(X)\|_2^2] < \infty$.

Hence, the strong duality applies here. Thus, $V(\beta)$ is convex and differentiable. The primal-dual problem is formulated as

$$\widehat{V}(\beta) = \in_Q \max_\lambda \mathcal{K}(Q) + \lambda(\mathtt{MI}(Q) - \beta).$$

The envelop theorem implies $\widehat{V}'(\beta) = \lambda^*(\beta) > 0$, where $\lambda^*(\beta)$ is the unique optimal dual variable (because $\mathcal{K}(Q) + \lambda(\mathtt{MI}(Q) - \beta)$ is strict convex in Q for $\lambda > 0$). Therefore, $\lambda^*(\beta)$ is non-decreasing.

Let $\widetilde{\beta}(\beta) = \beta - \mathtt{Gap}_d(Q(\beta)) < \beta$. Since the Gaussian noise $B_{\mathrm{PAC}}(\beta)$ satisfies $\mathtt{MI}_{\mathtt{PAC}}(B_{\mathrm{PAC}}(\beta)) = \widetilde{\beta}(\beta)$, we have

$$V_{\mathrm{PAC}}(\beta) = \mathcal{K}(B_{\mathrm{PAC}}(\beta)) \geq V(\widetilde{\beta}(\beta)).$$

Since $\widetilde{\beta}(\beta) < \beta$ and V is strictly increasing, $V(\widetilde{\beta}(\beta)) > V(\beta)$. Therefore, for all $\beta > 0$,

$$\Delta(\beta) \equiv V_{\mathrm{PAC}}(\beta) - V_{\mathrm{SR}}(\beta) > 0,$$

and $\lim_{\beta\to 0^+}\Delta(\beta) = 0$.

By Lemma 2 (stated and proved below) to $g(\beta) = V_{\mathrm{PAC}}(\beta)$ and $f(\beta) = V(\beta)$, we have $g'(\beta) > f'(\beta)$ for all $\beta > 0$. That is, $V'_{\mathrm{PAC}}(\beta) > V'_{\mathrm{SR}}(\beta)$. Thus, $\mathtt{Util}_\beta^{\mathrm{SR}} \geq \mathtt{Util}_\beta^{\mathrm{PAC}}$, with equality only for Gaussian $\mathcal{M}(X)$.

□

Lemma 2 (Height gap ⇒ slope gap)**.** *Let $g, f : (0,\infty) \to \mathbb{R}$ be differentiable, and assume f is convex. If $g(\beta) > f(\beta)$ for every $\beta > 0$ and $g(0) = f(0)$, then $g'(\beta) > f'(\beta)$ for every $\beta > 0$.*

Proof. Fix $\beta > 0$. For $h > 0$ small, $f(\beta + h) \geq f(\beta) + hf'(\beta)$ by convexity. Hence

$$\frac{g(\beta+h) - g(\beta)}{h} \ \geq\ f'(\beta) + \frac{g(\beta) - f(\beta)}{h}.$$

Sending $h \downarrow 0$ gives $g'(\beta) \geq f'(\beta)$. If equality held we would need $g(\beta) = f(\beta)$, contradicting the strict height gap. Hence $g'(\beta) > f'(\beta)$. □