

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

Tao Zhang

Washington University in St. Louis

Yevgeniy Vorobeychik

Washington University in St. Louis

## Abstract

The Probably Approximately Correct (PAC) Privacy framework [46] provides a powerful instance-based methodology to preserve privacy in complex data-driven systems. Existing PAC Privacy algorithms (we call them Auto-PAC) rely on a Gaussian mutual information upper bound. However, we show that the upper bound obtained by these algorithms is tight if and only if the perturbed mechanism output is jointly Gaussian with independent Gaussian noise. We propose two approaches for addressing this issue. First, we introduce two tractable post-processing methods for Auto-PAC, based on Donsker–Varadhan representation and sliced Wasserstein distances. However, the result still leaves "wasted" privacy budget. To address this issue more fundamentally, we introduce *Residual-PAC (R-PAC) Privacy*, an  $f$ -divergence-based measure to quantify privacy that remains after adversarial inference. To implement R-PAC Privacy in practice, we propose a Stackelberg Residual-PAC (SR-PAC) privatization mechanism, a game-theoretic framework that selects optimal noise distributions through convex bilevel optimization. Our approach achieves efficient privacy budget utilization for arbitrary data distributions and naturally composes when multiple mechanisms access the dataset. Through extensive experiments, we demonstrate that SR-PAC obtains consistently a better privacy-utility tradeoff than both PAC and differential privacy baselines.

## 1 Introduction

Machine-learning models power critical applications—from medical diagnosis to autonomous vehicles—yet their outputs can inadvertently expose sensitive training data. As pipelines grow in scale and complexity, practitioners need rigorous, scalable privacy guarantees that go beyond ad-hoc testing. Over the past two decades, formal privacy frameworks have proliferated. Differential Privacy (DP) [13] (and its variants such as Rényi DP [33]) delivers input-independent worst-case indistinguishability by bounding output shifts from single-record

changes. Alternative information-theoretic definitions, such as mutual-information DP [10], Fisher-information bounds [16, 21, 23], and Maximal Leakage [26, 39], provide complementary guarantees and offer alternative trade-offs between privacy and utility.

Nevertheless, provable privacy guarantees for modern data-processing algorithms remains a challenge. First, worst-case frameworks like DP require computing global sensitivity, which is generally NP-hard [47]. Moreover, computing the optimal privacy bound of DP under composition is, in general, a #P-complete [34]. In practice, finding the minimal noise needed to meet a target guarantee is intractable for most real-world algorithms, especially when the effect of each operation on privacy is unclear. On the other hand, empirical or simulation-based methods (e.g., testing resistance to membership inference [42]) address specific threats but lack rigorous, adversary-agnostic assurance. Bridging this gap requires a new, broadly applicable framework that can quantify and enforce privacy risk without relying on sensitivity.

A promising alternative has recently emerged: the Probably Approximately Correct (PAC) Privacy framework [46]. PAC Privacy shifts from indistinguishability-based guarantees to an operational notion that measures the *information-theoretic hardness* of reconstructing sensitive data. It is defined by an impossibility-of-inference guarantee for a chosen adversarial task and data prior, and the framework provides algorithms that enforce tractable mutual-information upper bounds to certify this guarantee. This approach enables automatic privatization via black-box simulation, and enjoys additive composition bounds and automatic privacy budget implementations for adaptive sequential compositions of mechanisms with arbitrary interdependencies. Notably, PAC Privacy often requires only  $O(1)$  noise magnitude to achieve its privacy guarantees—*independent of the output dimension*—whereas differential privacy’s worst-case, input-independent noise magnitude scales as  $\Theta(\sqrt{d})$  for a  $d$ -dimensional release.

However, existing PAC privacy algorithms are fundamentally conservative. In particular, we show (Proposition 1) that Auto-PAC achieves the designated privacy budget exactly if

and only if the perturbed mechanism output is jointly Gaussian with independent Gaussian noise, a restrictive condition rarely met in practice. Consequently, Auto-PAC will in general make inefficient use of the privacy budget.

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

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

In summary, our main contributions are as follows:

- We characterize the conservativeness of Auto-PAC [43, 46], showing that it arises from the gap between the Gaussian surrogate bound and the true non-Gaussian mutual information of the privatized mechanism.
- We propose two computationally tractable approaches to reduce this gap: one based on the Donsker–Varadhan representation (Theorem 3) and another based on the sliced Wasserstein distances (Theorem 4), both providing sample-efficient non-Gaussianity corrections.
- We propose a novel privacy framework, Residual-PAC (R-PAC), to quantify the portion of privacy that remains rather than the amount leaked. This offers a complementary perspective to PAC privacy, and enables efficient computation of tight privacy bounds.
- We present an automatic privatization algorithm, Stackelberg R-PAC (SR-PAC), to efficiently compute noise

distributions for a given privacy budget. SR-PAC algorithm achieves tight budget utilization, can operate with only black-box access via Monte Carlo simulation, and adaptively concentrates noise in privacy-sensitive directions while preserving task-relevant information.

## 1.1 Related Work

**Privacy Quantification Notions.** Quantitative notions of privacy leakage have been extensively studied across a variety of contexts, leading to mathematically rigorous frameworks 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. [13, 14] formalizing privacy loss through bounds on the distinguishability of outputs under neighboring datasets. Variants such as concentrated differential privacy (CDP) [5, 15], zero-concentrated DP (zCDP) [4], and Rényi differential privacy (RDP) [33] have further extended this framework by parameterizing privacy loss with different statistical divergences (e.g., Rényi divergence), thereby enhancing flexibility in privacy accounting, especially for compositions and adaptive mechanisms. 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, 10], with  $f$ -divergence and Fisher information offering finer-grained or context-specific metrics [16, 21, 23, 46]. 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 [1, 12, 18–20, 22, 30, 32, 40]. For example, Ghosh et al. [19] demonstrated that the geometric mechanism is universally optimal for differential privacy under certain loss-minimizing criteria in Bayesian settings, while Lebanon et al. [30] and Alghamdi et al. [1] studied utility-constrained optimization. Gupte et al. [22] 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.

**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) [24] and related GAN-based frameworks [8, 27, 35]—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. [41] 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

### 2.1 PAC Privacy

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

The PAC privacy framework [46] addresses this threat model and is formally defined as follows.

**Definition 1** (( $\delta, \rho, \mathcal{D}$ )-PAC Privacy [46]). *For a data processing mechanism  $\mathcal{M}$ , given some data distribution  $\mathcal{D}$ , and a measure function  $\rho(\cdot, \cdot)$ , we say  $\mathcal{M}$  satisfies ( $\delta, \rho, \mathcal{D}$ )-PAC Privacy if the following experiment is impossible:*

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

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  $|\tilde{X} - X|_2 \leq \epsilon$  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,  $\rho$  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  $\rho$  to encode a wide range of threat models and user-specified risk criteria. For example, in membership inference attacks [6],  $\rho(\tilde{X}, X) = 1$  may indicate that  $\tilde{X}$  successfully determines the presence of a target data point  $u_0$  in  $X$ . In reconstruction attacks [2], success may be defined by  $\rho(\tilde{X}, X) = 1$  if  $|\tilde{X} - X|_2 \leq 1$ , representing a close approximation of the original data.

Given the data distribution  $\mathcal{D}$  and the adversary's criterion  $\rho$ , the *optimal prior success rate* ( $1 - \delta_o^p$ ) is defined as the highest achievable success probability for the adversary without observing the output  $\mathcal{M}(X)$ :  $\delta_o^p = \inf_{\tilde{X}_0} \Pr_{X \sim \mathcal{D}} (\rho(\tilde{X}_0, X) \neq 1)$ . 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  $f : (0, +\infty) \rightarrow \mathbb{R}$  with  $f(1) = 0$ , extend  $f$  to  $t = 0$  by setting  $f(0) = \lim_{t \rightarrow 0^+} f(t)$  (in  $\mathbb{R} \cup \{+\infty, -\infty\}$ ). The  $f$ -divergence between two probability distributions  $P$  and  $Q$  over a common measurable space is:*

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

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

**Definition 3** (( $\Delta_f^\delta, \rho, \mathcal{D}$ ) PAC Advantage Privacy [46]). *A mechanism  $\mathcal{M}$  is termed ( $\Delta_f^\delta, \rho, \mathcal{D}$ ) PAC-advantage private if it is ( $\delta, \rho, \mathcal{D}$ ) PAC private and*

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

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

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

### 2.2 Automatic PAC Privatization Algorithms

PAC Privacy enables automatic privatization, which supports simulation-based implementation for arbitrary black-box mechanisms, without requiring the worst-case adversarial analysis, such as sensitivity computation. In this section, we present the main theorems and algorithms underlying automatic PAC privatization as introduced in [46] (hereafter "Auto-PAC") and the efficiency-improved version proposed in [43] (hereafter "Efficient-PAC"; algorithm details in Appendix A). We start by defining the *mutual information*.

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

$$\text{MI}(x; w) \equiv \mathcal{D}_{KL}(\mathbb{P}_{x,w} \| \mathbb{P}_x \otimes \mathbb{P}_w),$$

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

When the  $f$ -divergence in  $\Delta_f^\delta$  is instantiated as the KL divergence (denoted as  $\Delta_{KL}^\delta$ ), Theorem 1 of [46] shows

$$\Delta_{KL}^\delta \leq \text{MI}(X; \mathcal{M}(X)), \quad (1)$$

Thus, one can control the posterior advantage  $\Delta_{KL}^\delta$  by bounding the mutual information between private data and the released output.

Next, we introduce the Auto-PAC. Consider a deterministic data processing mechanism  $\mathcal{M} : X \rightarrow \mathbb{R}^d$ , 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  $B \sim \mathcal{N}(0, \Sigma_B)$ , where  $\Sigma_B$  is the covariance. For any deterministic mechanism  $\mathcal{M}$  and any Gaussian noise  $B$ , define the *Gaussian surrogate bound*

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

where  $\Sigma_{\mathcal{M}(X)}$  is the covariance of  $\mathcal{M}(X)$ .

**Theorem 1** (Theorem 3 of [46]). For an arbitrary deterministic mechanism  $\mathcal{M}$  and Gaussian noise  $B \sim \mathcal{N}(0, \Sigma_B)$ , the mutual information satisfies

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

Moreover, there exists  $\Sigma_B$  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  $\text{MI}(X; \mathcal{M}(X) + B) \leq \frac{1}{2}$ .

Theorem 1 establishes a simple upper bound on the mutual information with Gaussian noise perturbation. Choosing  $\Sigma_B$  to implement the Gaussian surrogate bound  $\text{LogDet}(\mathcal{M}(X), B) = \beta$  for a privacy budget  $\beta$  enables *anisotropic* noise as it estimates the eigenvectors of  $\mathcal{M}(X)$  to fit the instance-based noise to the geometry of the eigenspace of  $\mathcal{M}(X)$ . The result extends naturally to randomized mechanisms of the form  $\mathcal{M}(X, \theta)$ , where  $\theta$  is a random seed (Corollary 2 of [46]). Building on Theorem 1, an automatic PAC privatization algorithm (Auto-PAC) shown in Algorithm 1 is proposed by [46] to determine an appropriate Gaussian noise covariance  $\Sigma_B$  to ensure that  $\text{MI}(X; \mathcal{M}(X) + B) \leq \beta$  with confidence at least  $1 - \gamma$ . This is achieved using the user-specific security parameter  $c$ , privacy budget partitions  $v$  and  $\beta'$  such that  $\beta = v + \beta'$  (Theorem 4 of [46]). We refer to Algorithm 1 as  $(1 - \gamma)$ -Confidence Auto-PAC.

---

#### Algorithm 1 $(1 - \gamma)$ -Confidence Auto-PAC [46]

---

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

```

1: for  $k = 1, 2, \dots, m$  do
2:   Generate  $X^{(k)}$  from  $\mathcal{D}$ . Record  $y^{(k)} = \mathcal{M}(X^{(k)})$ .
3: end for
4: Calculate  $\hat{\mu} = \sum_{k=1}^m y^{(k)} / m$  and  $\hat{\Sigma} = \sum_{k=1}^m (y^{(k)} - \hat{\mu})(y^{(k)} - \hat{\mu})^T / m$ .
5: Apply SVD:  $\hat{\Sigma} = \hat{U} \hat{\Lambda} \hat{U}^T$ , where  $\hat{\Lambda}$  has eigenvalues  $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \dots \geq \hat{\lambda}_d$ .
6: Find  $j_0 = \arg \max_j \hat{\lambda}_j$  for  $\hat{\lambda}_j > c$ .
7: if  $\min_{1 \leq j \leq j_0, 1 \leq l \leq d} |\hat{\lambda}_j - \hat{\lambda}_l| > r\sqrt{dc} + 2c$  then
8:   for  $j = 1, 2, \dots, d$  do
9:     Set  $\lambda_{B,j} = \frac{2v}{\sqrt{\hat{\lambda}_j + 10cv/\beta'} \cdot (\sum_{j=1}^d \sqrt{\hat{\lambda}_j + 10cv/\beta'})}$ .
10:   end for
11:   Set  $\Sigma_B = \hat{U} \Lambda_B^{-1} \hat{U}^T$ .
12: else
13:   Set  $\Sigma_B = (\sum_{j=1}^d \hat{\lambda}_j + dc) / (2v) \cdot I_d$ .
14: end if
15: Output:  $\Sigma_B$ .
```

---

### 2.3 Differential Privacy

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

**Definition 5** ( $(\epsilon, \bar{\delta})$ -Differential Privacy [14]). A randomized mechanism  $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$  is said to be  $(\epsilon, \bar{\delta})$ -differentially private (DP), with  $\epsilon \geq 0$  and  $\bar{\delta} \in [0, 1]$ , if for any pair of adjacent datasets  $x, x'$ , and any measurable  $\mathcal{W} \subseteq \mathcal{Y}$ , it holds that  $\Pr[\mathcal{M}(x) \in \mathcal{W}] \leq e^\epsilon \Pr[\mathcal{M}(x') \in \mathcal{W}] + \bar{\delta}$ .

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

### 3 Characterizing The Gaussian Barrier of Automatic PAC Privatization

This section characterizes the utility of Auto-PAC by focusing on the conservativeness of the implemented mutual information bounds. To distinguish from Algorithm 1 ( $(1 - \gamma)$ -confidence Auto-PAC), we use Auto-PAC (algorithm) to refer



to the direct implementation of privacy budgets for the bound  $\text{LogDet}(\mathcal{M}(X), B)$  without a target conference level.

The Gaussian surrogate bound is conservative due to a nonzero *Gaussianity gap*, the discrepancy between the *true mutual information* and  $\text{LogDet}(\mathcal{M}(X), B)$  defined by (2):

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

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

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

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

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

Proposition 1 shows that the conservativeness of  $\mathcal{M}(X)$  in terms of  $\text{Gap}_d$  is equivalent to the KL divergence  $\text{D}_{\text{KL}}(P_{\mathcal{M}, B} \| \tilde{Q}_{\mathcal{M}})$ . Let  $\tilde{Z} \sim \tilde{Q}_{\mathcal{M}}$ . Then,  $\text{MI}(X; \tilde{Z}) = \text{LogDet}(\mathcal{M}(X), B)$ .

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

$$\inf_{B \sim Q} \mathbb{E}_Q[\|B\|_2^2] \quad \text{s.t.} \quad \text{MI}(X; \tilde{Z}) \leq \beta \text{ with } \tilde{Z} \sim \tilde{Q}_{\mathcal{M}}. \quad (5)$$

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

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

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

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

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

### 3.1 Mechanism Comparison in PAC Privacy

Definition 9 of [46] defines the optimal perturbation that tightly implements the privacy budget while maintaining optimal utility, where utility is captured by a loss function  $\mathcal{K}$ .

An optimal perturbation  $Q^*$  is a solution of the following optimization problem:

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

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

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

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

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

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

However, if Auto-PAC is used to solve the optimization problem (5), we have the conservative implementation of a given privacy budget. The next result shows that when  $\text{Gap}_d = \text{D}_{\text{KL}}(P_{\mathcal{M}, B} \| \tilde{Q}_{\mathcal{M}}) > 0$ , Auto-PAC does not, in general, maintain coherent ordering of PAC Privacy.

With a slight abuse of notation, for any mechanism  $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$ , let  $\text{Gap}_d(Q) = \text{D}_{\text{KL}}(P_{\mathcal{M}, B} \| \tilde{Q}_{\mathcal{M}})$  with  $B \sim Q$ .

**Theorem 2.** *Fix a mechanism  $\mathcal{M}$  and data distribution  $\mathcal{D}$ . Let  $\mathcal{Q}$  denote the collection of all zero-mean noise distributions under consideration, and let  $\text{I}_{\text{true}} : \mathcal{Q} \mapsto \mathbb{R}_{\geq 0}$  be the true mutual information functional; i.e.,  $\text{I}_{\text{true}}(Q) = \text{MI}(X; \mathcal{M}(X) + B)$  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  $G(\beta_2, \beta_1) \equiv \text{Gap}_d(Q^*(\beta_2)) - \text{Gap}_d(Q^*(\beta_1))$ . Then:*

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

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

Theorem 2 characterizes when Auto-PAC maintains coherent ordering of actual information leakage  $\text{I}_{\text{true}} = \beta - \text{Gap}_d$ . Increasing the budget from  $\beta_1$  to  $\beta_2$  permits extra leakage  $\beta_2 - \beta_1$  by using Auto-PAC, but part may be wasted if the mechanism output becomes more non-Gaussian. The wasted portion is  $G(\beta_2, \beta_1) = \text{Gap}_d(Q^*(\beta_2)) - \text{Gap}_d(Q^*(\beta_1))$ . If this waste exceeds the budget increase, then  $\text{I}_{\text{true}}$  decreases despite

a larger nominal budget, violating coherent ordering. This result cautions against comparing mechanisms using Auto-PAC solely by budgets, as identical budgets may yield different true leakages depending on their respective Gaussianity gaps.

### 3.2 Gap<sub>d</sub> Reduction via Non-Gaussianity Correction

In this section, we propose two approaches to reduce Gap<sub>d</sub> after a  $\mathcal{N}(0, \Sigma_B)$  is determined by Auto-PAC, enabling better estimation of the true mutual information to save privacy budgets. For any deterministic mechanism  $\mathcal{M}$  and Gaussian noise  $B \sim \mathcal{N}(0, \Sigma_B)$ , recall the Gaussian surrogate distribution  $\tilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$  in (4). Let  $D_Z = D_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}})$ . By Proposition 1, Gap<sub>d</sub> =  $D_Z$ .

For any estimator  $\hat{D}_Z$  of  $D_Z$ , define the improved mutual information estimate:

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

For  $0 \leq \hat{D}_Z \leq D_Z$ , we have

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

Thus, if we can obtain  $\hat{D}_Z$  satisfying  $0 \leq \hat{D}_Z \leq D_Z$  after Auto-PAC privatization, then for any  $\Sigma_B$  that ensures  $\text{LogDet}(\mathcal{M}(X), B) = \beta$ , we have  $\text{IMI}(\hat{D}_Z) = \beta - \hat{D}_Z$  as surrogate upper bound that is tighter than  $\text{LogDet}(\mathcal{M}(X), B)$ . Thus, we can have tighter privacy accounting post-hoc to the Auto-PAC privatization to save additional privacy budget, without requiring direct mutual information estimation.

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

**Definition 6** (Donsker–Varadhan (DV) Objective [11]). For probability measures  $P$  and  $Q$  on a common measurable space,

$$D_{\text{KL}}(P \| Q) = \sup_{f: \mathcal{Y} \rightarrow \mathbb{R}} \left\{ \mathbb{E}_P[f(Y)] - \log \mathbb{E}_Q[e^{f(Y)}] \right\},$$

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

**Definition 7** (Sliced Wasserstein Distances (SWD)). For  $p \geq 1$ , the  $p$ -Wasserstein distance between  $P$  and  $Q$  on  $\mathbb{R}^d$  is

$$W_p(P, Q) = \left( \inf_{\eta \in \hat{\Pi}(P, Q)} \mathbb{E}_{(X, Y) \sim \eta} [\|X - Y\|_2^p] \right)^{1/p},$$

where  $\hat{\Pi}(P, Q)$  is the set of couplings with marginals  $P$  and  $Q$ . The sliced  $p$ -Wasserstein distance averages 1-D Wasserstein distances over directions  $v$  on the unit sphere  $\mathbb{S}^{d-1}$ :

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

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

In addition, we define the following finite-sample lower-confidence estimator.

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

Let  $\Gamma_{\hat{\delta}} = \Gamma_{\hat{\delta}}(\mathcal{F}, n_P^{\text{val}}, n_Q^{\text{val}})$  be any valid uniform deviation bound satisfying, with probability at least  $1 - \hat{\delta}$ ,  $\sup_{f \in \mathcal{F}} \left| \hat{\mathcal{J}}(f; S_P^{\text{val}}, S_Q^{\text{val}}) - \mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_{\mathcal{M}}) \right| \leq \Gamma_{\hat{\delta}}$ , where  $\mathcal{J}(f; P, Q) = \mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f]$ . The finite-sample lower-confidence estimator of  $D_Z = D_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}})$  is

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

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

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

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

- (i)  $0 \leq \sup_f \hat{D}_Z(f) = D_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}})$ .
- (ii) For every  $f$ ,  $\hat{D}_Z(f) \leq D_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}}) \equiv D_Z$ .
- (iii) With probability at least  $1 - \hat{\delta}$  (over the independent validation splits in Definition 8),  $0 \leq \hat{D}_{\text{LCE}} \leq D_Z$ .

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

$$\hat{D}_Z \equiv \frac{1}{2\lambda_{\max}(\Sigma_Z)} \text{SW}_2^2(P_{\mathcal{M},B}, \tilde{Q}_{\mathcal{M}}).$$

Then  $0 \leq \hat{D}_Z \leq D_Z$ .

Theorems 3 and 4 lead to Corollary 1.

**Corollary 1.** Let  $\mathcal{M}: \mathcal{X} \mapsto \mathbb{R}^d$  be an arbitrary deterministic mechanism and  $B \sim \mathcal{N}(0, \Sigma_B)$  such that  $\text{LogDet}(\mathcal{M}(X), B) =$

$\beta$ . Under the assumptions of Theorems 3 and 4, the perturbed mechanism  $Z = \mathcal{M}(X) + B$  is PAC private with

$$\text{MI}(X; Z) \leq \beta - \widehat{D}_Z < \beta, \quad (8)$$

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

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

## 4 Residual-PAC Privacy

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

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

$$\text{IntP}_f(\mathcal{D}) = -D_f(\mathcal{D} \parallel \mathcal{R}),$$

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

**Examples of  $\mathcal{R}$ .** When  $\mathcal{X}$  is bounded,  $\mathcal{R}$  can be the uniform law  $\mathcal{U}$  on  $\mathcal{X}$ . However, on an unbounded  $\mathcal{X}$ , the uniform reference  $\mathcal{R} = \mathcal{U}$  has infinite volume  $\int_{\mathcal{X}} dx = \infty$ , potentially making  $\text{IntP}_f(\mathcal{D})$  vacuous or undefined. To avoid this, we instead require  $\mathcal{R}$  to satisfy  $D_f(\mathcal{D} \parallel \mathcal{R}) < \infty$ . For example, one can choose  $\mathcal{R}$  by: (i) truncated uniform on a large but bounded region containing  $\text{supp}(\mathcal{D})$ , (ii) maximum-entropy Gaussian matching known moments of  $\mathcal{D}$ , or (iii) smooth pullback of uniform on  $(0, 1)^d$  via bijection (e.g., component-wise sigmoid). Under any of these constructions,  $\mathcal{R}$  retains

the "random-guess" semantics yet has finite  $D_f(\mathcal{D} \parallel \mathcal{R})$ , ensuring  $\text{IntP}_f(\mathcal{D})$  remains meaningful even on unbounded  $\mathcal{X}$ . Please see Appendix D for a detailed discussion.

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

$$R_f^\delta \equiv \text{IntP}_f(\mathcal{D}) - D_f(\mathbf{1}_\delta \parallel \mathbf{1}_{\delta^p}),$$

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

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

$$\text{IntP}_f(\mathcal{D}) = R_f^\delta + \Delta_f^\delta. \quad (9)$$

This relationship provides a complete and interpretable quantification of privacy risk, distinguishing between the privacy that is lost and that which endures after information disclosure via  $\mathcal{M}(X)$ . Analogous to PAC Privacy, membership inference attacks (MIA) and R-PAC Membership Privacy can be instantiated from R-PAC Privacy. See Appendix B for detailed constructions.

### 4.1 Foundation of Residual-PAC Privacy

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

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

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

while for continuous  $X$ , the *differential entropy* is

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

**Conditional Entropy.** For jointly distributed random variables  $(X, W)$  where  $X$  is on alphabet  $\mathcal{X}$  and  $W$  is on alphabet  $\mathcal{W}$  (discrete case) or domain  $\mathcal{W}$  (continuous case), the *conditional entropy* of  $X$  given  $W$  is defined as

$$\mathcal{H}(X|W) = \sum_{w \in \mathcal{W}} P_W(w) \mathcal{H}(X|W = w)$$

in the discrete case and

$$h(X|W) = \int_{\mathcal{W}} f_W(w) h(X|W = w) dw$$

in the continuous case, where  $\mathcal{H}(X|W = w) = -\sum_{x \in \mathcal{X}} P_{X|W}(x|w) \log P_{X|W}(x|w)$  and  $h(X|W = w) = -\int_{\mathcal{X}} f_{X|W}(x|w) \log f_{X|W}(x|w) dx$ .

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

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

Consider any  $f$ -divergence  $D_f$ , Theorem 1 of [46] shows that the posterior advantage  $\Delta_f^\delta$  is bounded by the minimum  $f$ -divergence between the joint distribution of  $(X, \mathcal{M}(X))$ , denoted by  $P_{X, \mathcal{M}(X)}$ , and the product of the marginal distribution  $P_X$  and any auxiliary output distribution  $P_W$  independent of  $X$ :

$$\Delta_f^\delta \leq \inf_{P_W} D_f(P_{X, \mathcal{M}(X)} \| P_X \otimes P_W), \quad (10)$$

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

Thus, for any  $f$ -divergence  $D_f$ , inequality (10) implies that a mechanism  $\mathcal{M} : \mathcal{X} \rightarrow \mathcal{Y}$  satisfies  $(R_f^\delta, \rho, \mathcal{D})$  R-PAC Privacy if

$$R_f^\delta \geq \text{IntP}_f(\mathcal{D}) - \inf_{P_W} D_f(P_{X, \mathcal{M}(X)} \| P_X \otimes P_W). \quad (11)$$

Let  $R$  be a random variable distributed according to the reference distribution  $\mathcal{R}$  over  $\mathcal{X}$ .

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

$$R_f^\delta \geq \mathcal{H}(X|\mathcal{M}(X)) - V,$$

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

Corollary 2 follows from Theorem 1 of [46] and establishes that when  $\mathcal{H}(X)$  is finite, residual privacy  $R_f^\delta$  is lower bounded by  $\mathcal{H}(X|\mathcal{M}(X)) - V$ , where  $V$  is independent of both data distribution  $\mathcal{D}$  and mechanism  $\mathcal{M}$ . Since  $V$  is constant,  $R_f^\delta - V$  effectively provides a privacy quantification lower-bounded by conditional entropy  $\mathcal{H}(X|\mathcal{M}(X))$ . More generally, inequality (11) holds without requiring  $\mathcal{H}(X) < \infty$ , provided  $D_f(\mathcal{D} \| \mathcal{R}) < \infty$ .

## 4.2 Stackelberg Automatic Residual-PAC Privatization

In this section, we present our algorithms for automatic R-PAC privatization when the  $f$ -divergence is instantiated with

KL divergence, under which worst-case residual privacy is quantified by conditional entropy. For a utility loss function  $\mathcal{K}$ , we define the optimal perturbation problem for any R-PAC privacy budget  $\hat{\beta}$  as:

$$\inf_Q \mathbb{E}_{Q, \mathcal{M}, \mathcal{D}}[\mathcal{K}(B; \mathcal{M})] \quad \text{s.t.} \quad \mathcal{H}(X|\mathcal{M}(X) + B) \geq \hat{\beta}, B \sim Q. \quad (12)$$

When  $\mathcal{H}(X)$  is finite, by the definition of mutual information, any solution  $Q^*$  to problem (12) also solves (6) with PAC privacy budget  $\beta = \mathcal{H}(X) - \hat{\beta}$ . Given  $\text{MI}(X; \mathcal{M}(X) + B) = \mathcal{H}(X) - \mathcal{H}(X|\mathcal{M}(X) + B)$  with finite  $\mathcal{H}(X)$ , solving the optimal perturbation problem (12) with conditional entropy constraints presents the same computational challenges as (6).

To address this limitation, we present a novel automatic privatization approach for R-PAC privacy, termed *Stackelberg Automatic Residual-PAC Privatization (SR-PAC)*. Our approach is based on a Stackelberg game-theoretic characterization of the optimization (12). We show that SR-PAC achieves optimal perturbation without wasting privacy budget. Consequently, when  $\mathbb{E}_{Q, \mathcal{M}, \mathcal{D}}[\mathcal{K}(B; \mathcal{M})] = \mathbb{E}_Q[\|B\|_2^2]$ , SR-PAC can achieve superior utility performance compared to Auto-PAC and Efficient-PAC (Appendix A) for the same mutual information privacy budget.

---

### Algorithm 2 Monte Carlo SR-PAC

---

**Require:** Privacy budget  $\hat{\beta}$ , decoder family  $\Pi_\phi$ , perturbation rule family  $\Gamma_\lambda$ , utility loss  $\mathcal{K}(\cdot)$ , learning rates  $\eta_\phi, \eta_\lambda$ , penalty weight  $\sigma$ , iterations  $T_\lambda, T_\phi$ , batch size  $m$

- 1: Initialize parameters  $\lambda, \phi \sim \text{init}()$
- 2: **for**  $t = 1, \dots, T_\lambda$  **do**
- 3:   **if**  $t \bmod T_\phi = 0$  **then**
- 4:     **Update Decoder:**
- 5:     **for**  $i = 1, \dots, T_\phi$  **do**
- 6:       Sample  $\{(x_j, b_j, y_j)\}_{j=1}^m$  where  $x_j \sim \mathcal{D}$ ,  $b_j \sim Q_\lambda$ ,  
 $y_j = \mathcal{M}(x_j) + b_j$
- 7:        $\hat{W} = \frac{1}{m} \sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$
- 8:        $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \hat{W}$
- 9:     **end for**
- 10:   **end if**
- 11:   **Update Perturbation Rule:**
- 12:   Sample  $\{(x_j, b_j, y_j)\}_{j=1}^m$  where  $x_j \sim \mathcal{D}$ ,  $b_j \sim Q_\lambda$ ,  $y_j = \mathcal{M}(x_j) + b_j$
- 13:    $H_c = \frac{1}{m} \sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$
- 14:    $\mathcal{L}_\lambda = \frac{1}{m} \sum_{j=1}^m \mathcal{K}(b_j) + \sigma(H_c - \hat{\beta})^2$
- 15:    $\lambda \leftarrow \lambda - \eta_\lambda \nabla_\lambda \mathcal{L}_\lambda$
- 16: **end for**
- 17: **return** Optimal parameters  $(\lambda^*, \phi^*)$

---

Our SR-PAC algorithm recasts the optimal perturbation problem (12) as a Stackelberg game between a *Leader* (who chooses the perturbation rule  $Q$ ) and a *Follower* (who chooses the decoder attempting to infer  $X$  from  $Y$ ). Let  $\Gamma$  denote



a rich family of noise distributions. Let  $\Pi = \{\pi : \pi(\cdot|y) \in \Delta(X), y \in \mathcal{Y}\}$  denote a rich family of decoder distributions (e.g., all conditional density functions on  $X$  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

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

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

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

Therefore, a profile  $(Q^*, \pi^*)$  is a *Stackelberg equilibrium* if it satisfies

$$\begin{cases} Q^* \in \arg \inf_{Q \in \Gamma} \mathbb{E}[\mathcal{K}(B; \mathcal{M})], \text{ s.t. } W(Q, \pi^*(Q)) \geq \hat{\beta}, \\ \pi^*(Q) \in \arg \inf_{\pi \in \Pi} W(Q, \pi). \end{cases} \quad (13)$$

When we consider output perturbation and the utility loss  $\mathcal{K}$  is chosen such that  $Q \mapsto \mathbb{E}_{X \sim P_X, B \sim Q} [\mathcal{K}(B; \mathcal{M})]$  is convex in  $Q$ , the problem (13) is convex in both  $Q$  and  $\pi$ . Specifically, for each fixed perturbation rule  $Q$ , the map  $\pi \mapsto W(Q, \pi)$  is a convex function of  $\pi$ . Similarly, for each fixed decoder  $\pi$ , 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  $\pi$  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  $\pi$  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  $Q \mapsto \mathbb{E}_{X \sim P_X, B \sim Q} [\mathcal{K}(B; \mathcal{M})]$  over this set remains a convex program in  $Q$ . Meanwhile, the Follower's problem  $\inf_{\pi \in \Pi} W(Q, \pi)$  is convex in  $\pi$  for any fixed  $Q$ . Thus, the Stackelberg game reduces to a single-level convex optimization in  $Q$ , with the inner decoder problem convex in  $\pi$ .

Proposition 5 shows that the Stackelberg equilibrium perturbation rule solves (12).

**Proposition 5.** *Let  $(Q^*, \pi^*)$  be a Stackelberg equilibrium satisfying (13) for any  $\hat{\beta}$ . Then,  $Q^*$  solves (12) with privacy budget  $\hat{\beta}$ . In addition, in any Stackelberg equilibrium  $(Q^*, \pi^*)$ ,  $\pi^* = \pi^*(Q^*)$  is unique.*

Algorithm 2 provides a Monte-Carlo-based approach to solve the Stackelberg equilibrium (13). 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 privacy budget.

## 5 Properties of SR-PAC Privatization

This section presents some important properties of SR-PAC.

### 5.1 Anisotropic Noise Perturbation

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

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

- (i) *Every  $Q \in \Gamma$  is log-concave.*
- (ii) *For any orthonormal direction  $w \in \mathbb{R}^d$ ,  $\langle \mathcal{M}(X), w \rangle$  is non-degenerate.*
- (iii) *The utility function  $\mathcal{K}$  is radial (depends only on  $\|B\|_2$ ) and strictly convex in the eigenvalues of covariance matrix  $\Sigma_Q$  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  $\Gamma$  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 6.** *Under Assumption 1, any Stackelberg-optimal perturbation rule  $Q^*$  is anisotropic. That is, its covariance matrix  $\Sigma_{Q^*}$  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  $\Sigma_{Q^*}$ .

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

## 5.2 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  $\Pi_{\text{task}}$  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  $S_{\text{lab}} := \text{span}\{e_\ell - e_j : j \neq \ell\}$ , where lab means "label". Let  $\Pi_{\text{lab}}$  be the projector onto  $S_{\text{lab}}$ . The analysis for a general  $S_{\text{task}}$  is identical after replacing lab by task.

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

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

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

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

- (i) Let  $\mathcal{N}(0, \Sigma_{\text{PAC}})$  be the Gaussian noise distribution used by the Auto-PAC such that  $\text{LogDet}(Z, B_{\text{PAC}}) = \beta$ . If  $Z$  is non-Gaussian, then  $\mathbb{E}_{Q^*}[\|B\|_2^2] < \mathbb{E}[\|B_{\text{PAC}}\|_2^2]$ .
- (ii) Suppose  $\sup_{v \in S_{\text{lab}}, \|v\|=1} g(v) < \inf_{w \perp S_{\text{lab}}, \|w\|=1} g(w)$ . Let  $\beta_{\text{lab}} \equiv \frac{1}{2} \int_{w \perp S_{\text{lab}}} g(w) d\sigma_w^2$  be the maximal MI reduction achievable with noise orthogonal to  $S_{\text{lab}}$ . Then, for every  $\beta \leq \beta_{\text{lab}}$ , we have  $\Pi_{\text{lab}} B^* = 0$  a.s.,  $\arg \max_i (Z_i + B_i^*) = \hat{y}$  a.s.

In Proposition 7, part (i) shows that SR-PAC always uses strictly less noise magnitude than any Auto-PAC (regardless of how anisotropic the Auto-PAC noise covariance may be) because Auto-PAC treats  $Z$  as Gaussian and thus overestimates the required variance when  $Z$  is non-Gaussian. Part (ii) demonstrates that, under the natural ordering of directional sensitivities, SR-PAC allocates its noise budget exclusively in directions orthogonal to the label sub-space until a critical threshold  $\beta_{\text{lab}}$  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.

## 5.3 Sensitivity to $\beta$

Sensitivity to the privacy parameter  $\beta$  is crucial for predictable and accurate control of privacy-utility trade-off. Let  $\text{Priv}_\beta$  and  $\text{Util}_\beta$ , respectively, denote the sensitivities of privacy and utility (for certain measures). If  $\text{Priv}_\beta = 1$ , then any infinitesimal increase  $\Delta\beta$  in the privacy budget raises the true mutual

information  $\text{MI}(X; Y)$  by exactly  $\Delta\beta$ . Thus, no part of the privacy budget is "wasted" or "over-consumed". By contrast, if  $\text{Priv}_\beta < 1$ , then increasing  $\beta$  may force additional noise without achieving the full allowed leakage; and if  $\text{Priv}_\beta > 1$ , even a small increase in  $\beta$  could exceed the allowed privacy. Similarly, if  $\text{Util}_\beta$  is high, then an infinitesimal increase  $\Delta\beta$  in the privacy budget yields a large improvement in utility; if  $\text{Util}_\beta$  is low, the same increase yields a small improvement, indicating inefficient conversion of the privacy budget into utility gains.

Let

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

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

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

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

Theorem 5 proves that SR-PAC with arbitrary noise distributions achieves: (i) *Exact leakage-budget alignment* ( $\text{Priv}_\beta^{\text{SR}} = 1$ ), (ii) *Stricter utility decay* for Auto-PAC ( $\text{Util}_\beta^{\text{SR}} \geq \text{Util}_\beta^{\text{PAC}}$ ). This holds unconditionally for non-Gaussian  $\mathcal{M}(X)$  under privacy tightening (i.e.,  $\beta$  decreasing).

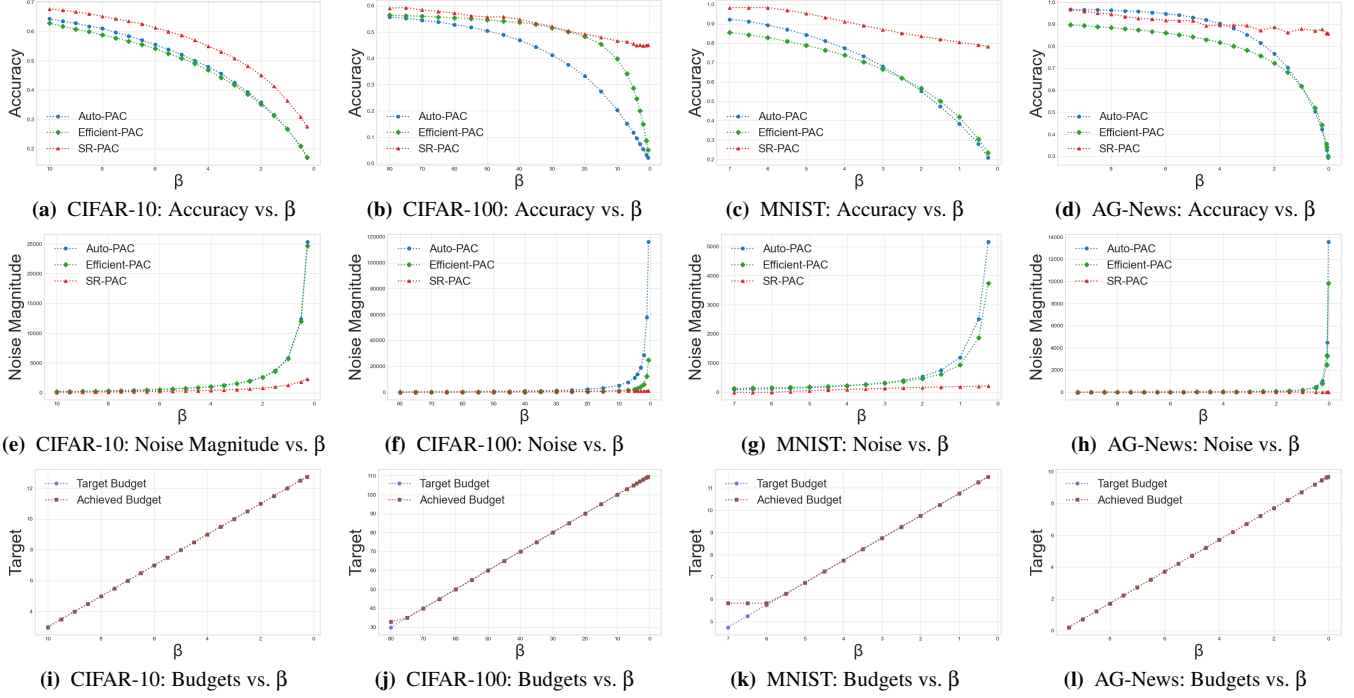
**Corollary 3.** In addition to the setting of Theorem 5, assume

$$\epsilon_{\text{cal}}(\beta) \in [0, \text{Gap}_d(\hat{Q}(\beta))], \quad \eta_{\text{opt}}(\beta) \in [0, V_{\text{PAC}}(\beta) - V_{\text{SR}}(\beta)).$$

Then, (i)  $|\text{Priv}_\beta^{\text{SR}} - 1| \leq |\epsilon'_{\text{cal}}(\beta)|$ ; (ii)  $\text{Util}_\beta^{\text{SR}} \geq \text{Util}_\beta^{\text{PAC}}$ , with equality only for Gaussian  $\mathcal{M}(X)$ .

## 5.4 Composition

Graceful composition properties in privacy definitions like DP enable quantifiable privacy risk across multiple operations on datasets, allowing modular system design where components maintain local privacy-utility trade-offs while preserving global privacy guarantees. Consider  $k$  mechanisms  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k$ , 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  $\vec{\mathcal{M}}(\cdot, \vec{\Theta}) : \mathcal{X} \mapsto \vec{\mathcal{Y}}$  is defined as

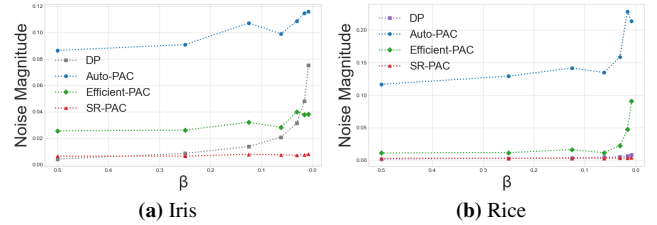


**Figure 1:** Empirical comparisons of SR-PAC, Auto-PAC (Algorithm 1), and Efficient-PAC (Algorithm 3) on CIFAR-10, CIFAR-100, MNIST, and AG-News as  $\beta$  varies. Each column corresponds to one dataset; within each column, the three panels report (top) classification accuracy of the perturbed model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

$\vec{\mathcal{M}}(X, \vec{\theta}) = (\mathcal{M}_1(X, \theta_1), \dots, \mathcal{M}_k(X, \theta_k))$ . 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  $\beta_i$ , then  $\vec{\mathcal{M}}$  has bound  $\sum_{i=1}^k \beta_i$ . R-PAC Privacy also enjoys additive composition with respect to conditional entropy bounds. Suppose each mechanism  $\mathcal{M}_i$  is R-PAC private with conditional entropy lower bound  $\hat{\beta}_i$ . By 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 Xiao et al. (2023), the composition  $\vec{\mathcal{M}}(X, \vec{\theta})$  is PAC private with total mutual information upper bounded by  $\sum_{i=1}^k (\mathcal{H}(X) - \hat{\beta}_i)$ . Equivalently, the composition  $\vec{\mathcal{M}}(X, \vec{\theta})$  is R-PAC private with overall conditional entropy lower bounded by  $\sum_{i=1}^k \hat{\beta}_i - (k-1)\mathcal{H}(X)$ .

However, this additive composition property for mutual information yields conservative aggregated privacy bounds, and utility degradation compounds when each mechanism  $\mathcal{M}_i$  uses conservative privacy budgets  $\beta_i$ . To address this limitation, we employ an optimization-based approach within the SR-PAC framework to compute tighter conditional entropy bounds. Consider  $k$  mechanisms  $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_k$  privatized by distributions  $Q_i$  to satisfy R-PAC privacy with bounds  $\hat{\beta}_i$ . The Leader designs these privatizations  $\{Q_i\}_{i=1}^k$ , while the Follower finds the optimal decoder for the joint composition



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

$$\vec{\mathcal{M}}(X, \vec{\theta}) = (\mathcal{M}_1(X), \dots, \mathcal{M}_k(X)):$$

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

This game-theoretic formulation allows for tighter privacy-utility trade-offs in composed systems by optimizing the joint privatization strategy.

## 6 Experiments

We conduct two sets of experiments to evaluate our approach. First, we compare SR-PAC against Auto-PAC and Efficient-PAC (Appendix A) using CIFAR-10 [29], CIFAR-100 [29],

MNIST [31], and AG-News [48] datasets, with results presented in Section 6.1. We use (R-)PAC to refer to the family of SR-PAC, Auto-PAC, and Efficient-PAC. Second, we extend this comparison to include DP by equalizing optimal posterior success rates of membership inference (Appendix C.2) across all four methods (SR-PAC, Auto-PAC, Efficient-PAC, and DP), making their privacy budgets comparable. For this comparison, we use Iris [17] and Rice [9] datasets, with results shown in Section 6.2. All experiments focus on output perturbation. Appendix S provides more details of the experiments.

**CIFAR-10 and Base Classifier.** The CIFAR-10 dataset comprises 50,000 training and 10,000 testing color images (each  $32 \times 32$  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. An unperturbed classifier is a convolutional neural network that consists of two convolutional blocks—each block is  $\text{Conv} \rightarrow \text{ReLU} \rightarrow \text{MaxPool}$  (kernel  $2 \times 2$ ) 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 classifier achieves  $0.7181 \pm 0.0088$  accuracy.

**CIFAR-100 and Base Classifier.** 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. The unperturbed classifier is a deep convolutional neural network with three convolutional “blocks.” Each block consists of two  $3 \times 3$  convolutions (with BatchNorm and ReLU after each), followed by a  $2 \times 2$  max-pool, which sequentially maps inputs from  $32 \times 32 \rightarrow 16 \times 16 \rightarrow 8 \times 8 \rightarrow 4 \times 4$ , with channel widths increasing from  $3 \rightarrow 64 \rightarrow 128 \rightarrow 256$ . After flattening the resulting  $256 \times 4 \times 4$  feature map into a 4096-dimensional vector, a three-layer MLP head ( $4096 \rightarrow 512 \rightarrow 256 \rightarrow 100$ ) 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 classifier achieves  $0.5914 \pm 0.0090$  accuracy.

**MNIST dataset and Base Classifier.** The MNIST dataset comprises 60,000 training and 10,000 test grayscale images of handwritten digits (0–9). Each image is  $28 \times 28$  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. The unperturbed classifier is a simple CNN consisting of two convolutional blocks—each block is  $\text{Conv2d} \rightarrow \text{BatchNorm} \rightarrow \text{ReLU} \rightarrow \text{MaxPool}$  ( $2 \times 2$ ), with channel widths  $1 \rightarrow 32 \rightarrow 64$ —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  $28 \times 28$  image is mapped to a 10-dimensional logit vector, and the predicted label is given by the index of the largest logit. The unperturbed classifier achieves 0.9837 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 ( $300 \rightarrow 256$  with ReLU and 0.3 dropout, then  $256 \rightarrow 4$ ), yielding a 4-dimensional logit vector, and at inference the predicted label is the index of the largest logit. The unperturbed mechanism achieves 0.9705 accuracy.

Recall that  $\beta$  upper-bounds  $\text{MI}(X; \mathcal{M}(X) + B)$ , and SR-PAC enforces the equivalent constraint  $\mathcal{H}(X | \mathcal{M}(X) + B) \geq \hat{\beta} = \mathcal{H}(X) - \beta$ . Although  $\mathcal{H}(X)$  is unknown, we estimate for the purpose of evaluation to verify the tightness of the privacy bounds. Let  $\text{MI}_0 = \text{MI}(X; \mathcal{M}(X))$ . By data processing,  $\text{MI}(X; \mathcal{M}(X) + B) \leq \text{MI}_0$  for any independent  $B$ , so the feasible budgets are  $0 < \beta \leq \text{MI}_0$  and this interval is common to Auto-PAC, Efficient-PAC, and SR-PAC. At  $\beta = \text{MI}_0$  the optimal choice is  $B = 0$ , and all methods coincide at the noiseless accuracy. This shared endpoint and feasible domain ensure that comparisons at a common target  $\beta$  are well-defined even without the exact  $\mathcal{H}(X)$ ; moreover, reparameterizing by achieved mutual information preserves the endpoint and domain, and—together with the small budget errors observed in panels (i–l)—does not affect our empirical ordering. Under additive  $\ell_2$  output noise, the ordering by total noise magnitude  $\mathbb{E}[\|B\|_2^2]$  coincides with the ordering by accuracy, consistent with the  $\ell_2$ -based behavior reported in prior work; hence the accuracy– and noise–vs.– $\beta$  panels convey the same conclusion in our experiments.

## 6.1 (R-)PAC Comparison

For each dataset and its pretrained base classifier  $\mathcal{M}$ , we plot (1) the test accuracy of the perturbed model as a function of  $\beta$ , (2) the average noise magnitude  $\mathbb{E}[\|B\|_2^2]$  required to achieve each  $\beta$ , and (3) SR-PAC’s ability to hit the target



budget  $\hat{\mathcal{H}}(X) - \beta$  (where  $\hat{\mathcal{H}}(X)$  is our entropy estimate).

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

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

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

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

## 6.2 Comparison with Differential Privacy

We calibrate DP and (R-)PAC to the same (optimal) posterior success rate for membership inference attacks, then compare their utility in terms of noise magnitudes (i.e.,  $\ell_2$ -norm of the difference between original and perturbed outputs). The base mechanism is a mean estimator. Appendix C.2 provides the conversions between (optimal) posterior success rates, DP parameters (DP→posterior mapping), and mutual information budgets (MI→posterior mapping). Concretely, for DP we select  $(\epsilon, \delta)$  that yields the target posterior bound via the DP→posterior mapping, and for (R-)PAC we choose  $\beta$  that yields the same posterior via the MI→posterior mapping; with subsampling rate  $r = 0.5$  we have prior  $p = 0.5$ . In each

trial, we construct a membership vector  $m \in \{0, 1\}^P$  by i.i.d. Bernoulli(0.5) draws, so the member count  $S = \sum_i m_i$  is random. The released statistic is the mean of the data. We follow similar treatments for DP as Section 6.3 of [43]: the DP baseline clips each row in  $\ell_2$  to radius  $C$ , adds calibrated Gaussian noise to the clipped sum, and divides by  $S$  to produce the privatized mean; (R-)PAC injects additive output noise calibrated to the target  $\beta$ . We report the average noise magnitude  $\mathbb{E}[\|B\|_2^2]$  at matched posterior success rates; in our output-perturbed mean setting, this quantity **equals** the expected squared  $\ell_2$  error of the released statistic (i.e., the  $\ell_2$  accuracy metric we use). Hence the ordering by  $\mathbb{E}[\|B\|_2^2]$  is identical to the ordering by  $\ell_2$  accuracy. Qualitative DP–(R-)PAC relations are discussed in Appendix C.2.

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

Auto-PAC and Efficient-PAC do allocate *anisotropic* noise, but their shapes are task-agnostic and depend only on second-order structure (the empirical covariance and its spectrum), via covariance scaling (Auto-PAC) or eigen-allocation (Efficient-PAC). In small-sample regimes such as Iris and Rice, the covariance spectrum is noisy and can be ill-conditioned. These moment-based rules propagate that instability into the noise design and calibration, leading to conservative (over-noisy) implementations—especially under tight budgets (small  $\beta$ ). By contrast, SR-PAC enforces the conditional-entropy budget directly, yielding tighter budget implementation and lower required noise. Empirically (Figure 2), SR-PAC achieves smaller average noise magnitudes across  $\beta$  and exhibits smoother scaling, indicating greater stability than Auto-PAC and Efficient-PAC on these small-sample tasks.

## 7 Conclusion

In this work, we introduced Residual-PAC Privacy, an enhanced framework that quantifies privacy guarantees beyond Gaussian assumptions while overcoming the conservativeness of prior PAC-Privacy methods. Our Stackelberg Residual-PAC (SR-PAC) approach casts the privacy-utility trade-off as a convex Stackelberg optimization problem, fully leveraging available privacy budgets and automatically calibrating anisotropic noise distributions tailored to specific data and mechanisms. Extensive experiments demonstrate that SR-PAC consistently achieves tighter privacy guarantees and higher utility than existing approaches, providing a rigorous yet practical foundation for scalable privacy assurance in complex applications.

## References

- [1] Wael Alghamdi, Shahab Asoodeh, Flavio P Calmon, Oliver Kosut, Lalitha Sankar, and Fei Wei. Cactus mechanisms: Optimal differential privacy mechanisms in the large-composition regime. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1838–1843, 2022.
- [2] Borja Balle, Giovanni Cherubin, and Jamie Hayes. Reconstructing training data with informed adversaries. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1138–1156. IEEE, 2022.
- [3] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.
- [4] Mark Bun, Cynthia Dwork, Guy N Rothblum, and Thomas Steinke. Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 74–86, 2018.
- [5] Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of cryptography conference*, pages 635–658. Springer, 2016.
- [6] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914. IEEE, 2022.
- [7] Konstantinos Chatzikokolakis, Tom Chothia, and Apratim Guha. Statistical measurement of information leakage. In *International Conference on Tools and Algorithms for the Construction and Analysis of Systems*, pages 390–404. Springer, 2010.
- [8] Xiao Chen, Peter Kairouz, and Ram Rajagopal. Understanding compressive adversarial privacy. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6824–6831. IEEE, 2018.
- [9] Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.
- [10] Paul Cuff and Lanqing Yu. Differential privacy as a mutual information constraint. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 43–54, 2016.
- [11] Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time, i. *Communications on pure and applied mathematics*, 28(1):1–47, 1975.
- [12] Flávio du Pin Calmon and Nadia Fawaz. Privacy against statistical inference. In *2012 50th annual Allerton conference on communication, control, and computing (Allerton)*, pages 1401–1408. IEEE, 2012.
- [13] Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pages 1–12, 2006.
- [14] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pages 265–284. Springer, 2006.
- [15] Cynthia Dwork and Guy N Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.
- [16] Farhad Farokhi and Henrik Sandberg. Fisher information as a measure of privacy: Preserving privacy of households with smart meters using batteries. *IEEE Transactions on Smart Grid*, 9(5):4726–4734, 2017.
- [17] Ronald Aylmer Fisher. Iris. uci machine learning repository. DOI: <https://doi.org/10.24432/C56C76>, 1988.
- [18] Quan Geng, Wei Ding, Ruiqi Guo, and Sanjiv Kumar. Tight analysis of privacy and utility tradeoff in approximate differential privacy. In *International Conference on Artificial Intelligence and Statistics*, pages 89–99. PMLR, 2020.
- [19] Arpita Ghosh, Tim Roughgarden, and Mukund Sundararajan. Universally utility-maximizing privacy mechanisms. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 351–360, 2009.
- [20] Jasper Goseling and Milan Lopuhaä-Zwakenberg. Robust optimization for local differential privacy. In *2022 IEEE International Symposium on Information Theory (ISIT)*, pages 1629–1634. IEEE, 2022.
- [21] Chuan Guo, Brian Karrer, Kamalika Chaudhuri, and Laurens van der Maaten. Bounding training data reconstruction in private (deep) learning. In *International Conference on Machine Learning*, pages 8056–8071. PMLR, 2022.
- [22] Mangesh Gupte and Mukund Sundararajan. Universally optimal privacy mechanisms for minimax agents. In *Proceedings of the twenty-ninth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 135–146, 2010.

- [23] Awni Hannun, Chuan Guo, and Laurens van der Maaten. Measuring data leakage in machine-learning models with fisher information. In *Uncertainty in Artificial Intelligence*, pages 760–770. PMLR, 2021.
- [24] Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Generative adversarial privacy. *arXiv preprint arXiv:1807.05306*, 2018.
- [25] Thomas Humphries, Simon Oya, Lindsey Tulloch, Matthew Rafuse, Ian Goldberg, Urs Hengartner, and Florian Kerschbaum. Investigating membership inference attacks under data dependencies. In *2023 IEEE 36th Computer Security Foundations Symposium (CSF)*, pages 473–488. IEEE, 2023.
- [26] Ibrahim Issa, Aaron B Wagner, and Sudeep Kamath. An operational approach to information leakage. *IEEE Transactions on Information Theory*, 66(3):1625–1657, 2019.
- [27] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*, 2018.
- [28] Peter Kairouz, Sewoong Oh, and Pramod Viswanath. The composition theorem for differential privacy. In *International conference on machine learning*, pages 1376–1385. PMLR, 2015.
- [29] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.
- [30] Guy Lebanon, Monica Scannapieco, Mohamed Fouad, and Elisa Bertino. Beyond k-anonymity: A decision theoretic framework for assessing privacy risk. *Transactions on Data Privacy*, 2009.
- [31] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]*. Available: <http://yann.lecun.com/exdb/mnist>, 2, 2010.
- [32] Milan Lopuhaä-Zwakenberg and Jasper Goseling. Mechanisms for robust local differential privacy. *Entropy*, 26(3):233, 2024.
- [33] Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*, pages 263–275. IEEE, 2017.
- [34] Jack Murtagh and Salil Vadhan. The complexity of computing the optimal composition of differential privacy. In *Theory of Cryptography Conference*, pages 157–175. Springer, 2015.
- [35] Milad Nasr, Reza Shokri, and Amir Houmansadr. Machine learning with membership privacy using adversarial regularization. In *Proceedings of the 2018 ACM SIGSAC conference on computer and communications security*, pages 634–646, 2018.
- [36] Felix Otto and Cédric Villani. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis*, 173(2):361–400, 2000.
- [37] Daniel P Palomar and Sergio Verdú. Gradient of mutual information in linear vector gaussian channels. *IEEE Transactions on Information Theory*, 52(1):141–154, 2005.
- [38] Sangwoo Park, Erchin Serpedin, and Khalid Qaraqe. On the equivalence between stein and de bruijn identities. *IEEE Transactions on Information Theory*, 58(12):7045–7067, 2012.
- [39] Sara Saeidian, Giulia Cervia, Tobias J Oechtering, and Mikael Skoglund. Pointwise maximal leakage. *IEEE Transactions on Information Theory*, 69(12):8054–8080, 2023.
- [40] Lalitha Sankar, S Raj Rajagopalan, and H Vincent Poor. Utility-privacy tradeoffs in databases: An information-theoretic approach. *IEEE Transactions on Information Forensics and Security*, 8(6):838–852, 2013.
- [41] Aras Selvi, Huikang Liu, and Wolfram Wiesemann. Differential privacy via distributionally robust optimization. *Operations Research*, 2025.
- [42] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.
- [43] Mayuri Sridhar, Hanshen Xiao, and Srinivas Devadas. Pac-private algorithms. *Cryptology ePrint Archive*, 2024.
- [44] Aart J Stam. Some inequalities satisfied by the quantities of information of fisher and shannon. *Information and Control*, 2(2):101–112, 1959.
- [45] Michel Talagrand. Transportation cost for gaussian and other product measures. *Geometric & Functional Analysis GAFA*, 6(3):587–600, 1996.
- [46] Hanshen Xiao and Srinivas Devadas. Pac privacy: Automatic privacy measurement and control of data processing. In *Annual International Cryptology Conference*, pages 611–644. Springer, 2023.
- [47] Xiaokui Xiao and Yufei Tao. Output perturbation with query relaxation. *Proceedings of the VLDB Endowment*, 1(1):857–869, 2008.

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

## A Automatic Efficient PAC Privatization

PAC privacy (Auto-PAC, Algorithm 1) provides a framework for measuring privacy risk through simulation-based proofs that bound the mutual information between inputs and outputs of black-box algorithms. While this approach offers rigorous privacy guarantees without requiring white-box algorithm modifications, the original implementation faced computational and practical challenges. The initial algorithm required computing the full covariance matrix and performing Singular Value Decomposition (SVD) across the entire output dimension, which becomes prohibitively expensive for high-dimensional outputs. Additionally, black-box privacy mechanisms suffer from output instability caused by random seeds, arbitrary encodings, or non-deterministic implementations, leading to inconsistent noise calibration and suboptimal utility.

Recent work by Sridhar et al. [43] addresses these limitations through Efficient-PAC (Algorithm 3), which introduces two key improvements. First, they develop an anisotropic noise calibration scheme that avoids full covariance estimation by projecting mechanism outputs onto a unitary basis and estimating only per-direction variances. This leads to a more scalable and sample-efficient algorithm while maintaining rigorous mutual information guarantees. Second, they propose methods for reducing output instability through regularization and canonicalization techniques, 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 6 establishes the privacy guarantee of Efficient-PAC.

**Theorem 6** (Theorem 1 of [43]). *Let  $\mathcal{M} : \mathcal{X} \rightarrow \mathbb{R}^d$  be a deterministic mechanism, and let  $A \in \mathbb{R}^{d \times d}$  be a unitary projection matrix. Let  $\sigma \in \mathbb{R}^d$  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  $\Sigma_B = \text{diag}(e_1, \dots, e_d)$ , 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  $\text{MI}(X; \mathcal{M}(X) + B) \leq \beta$ .*

### A.1 Auto-PAC vs. Efficient-PAC: Conservativeness

Efficient-PAC induces additional conservativeness relative to Auto-PAC. When Efficient-PAC enforces  $\text{MI}(X; \mathcal{M}(X) +$

---

### Algorithm 3 Efficient-PAC [43]

---

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

- 1: Initialize  $m \leftarrow 1$ ,  $\sigma_0 \leftarrow \text{null}$ ,  $\mathbf{G} \leftarrow \text{null}$
- 2: **while**  $m \leq 2$  or  $f_\tau(\sigma_{m-1}, \sigma_m) \geq \tau$  **do**
- 3:   Sample  $X_m \sim \mathcal{D}$ , compute  $y_m \leftarrow \mathcal{M}(X_m)$
- 4:   Set  $g_m \leftarrow [y_m \cdot A_1, \dots, y_m \cdot A_d]$ , append to  $\mathbf{G}$
- 5:   Set  $\sigma_m[k]$  to empirical variance of column  $k$  in  $\mathbf{G}$ , increment  $m \leftarrow m + 1$
- 6: **end while**
- 7: **for**  $i = 1$  to  $d$  **do**
- 8:   Set  $e_i \leftarrow \frac{\sqrt{\sigma_m[i]}}{2\beta} \sum_{j=1}^d \sqrt{\sigma_m[j]}$
- 9: **end for**
- 10: **return**  $\Sigma_B$  with  $\Sigma_B[i][i] = e_i$

---

$B) \leq \beta$ , the proof of Theorem 6 in [43] (Theorem 1) yields

$$\begin{aligned}
\text{MI}(X; \mathcal{M}(X) + B) &= \text{MI}(X; \mathcal{M}(X) \cdot A + B \cdot A) \\
&\leq \frac{1}{2} \log \det \left( I_d + \Sigma_{\mathcal{M}(X) \cdot A} \Sigma_B^{-1} \right) \\
&\leq \frac{1}{2} \log \det \left( I_d + \text{diag}(\Sigma_{\mathcal{M}(X) \cdot A}) \Sigma_B^{-1} \right) \\
&= \frac{1}{2} \log \prod_i \left( 1 + \frac{\sigma_i}{e_i} \right) \\
&= \frac{1}{2} \sum_i \log \left( 1 + \frac{\sigma_i}{e_i} \right) \\
&\leq \frac{1}{2} \sum_i \frac{\sigma_i}{e_i} \\
&= \beta,
\end{aligned}$$

where  $\sigma_i = [\text{diag}(\Sigma_{\mathcal{M}(X) \cdot A})]_i$  and  $\Sigma_B = \text{diag}(e_1, \dots, e_d)$ . The second inequality is Hadamard's inequality (tight only if  $\Sigma_{\mathcal{M}(X) \cdot A}$  is diagonal in the chosen basis), and the last inequality uses  $\log(1+x) \leq x$  (tight only at  $x=0$ ). Minimizing  $\sum_i e_i$  under the linearized constraint  $\frac{1}{2} \sum_i \sigma_i / e_i = \beta$  gives the closed form  $e_i = \frac{\sum_j \sqrt{\sigma_j}}{2\beta} \sqrt{\sigma_i}$ , so that  $\sum_{i=1}^d \frac{\sigma_i}{2e_i} = \beta$ . Thus, Efficient-PAC is *weakly more conservative* than Auto-PAC, which (approximately) works in the eigenbasis of  $\Sigma_{\mathcal{M}(X)}$  and avoids the Hadamard slack.

Moreover, since  $B \cdot A$  is constructed with covariance  $\Sigma_{B \cdot A} = A^\top \Sigma_B A$  and  $\Sigma_{\mathcal{M}(X) \cdot A} = A^\top \Sigma_{\mathcal{M}(X)} A$ , the exact log-det term is basis-invariant under joint congruence:

$$\begin{aligned}
\frac{1}{2} \log \det \left( I_d + \Sigma_{\mathcal{M}(X) \cdot A} \Sigma_{B \cdot A}^{-1} \right) &= \frac{1}{2} \log \frac{\det(\Sigma_{B \cdot A} + \Sigma_{\mathcal{M}(X) \cdot A})}{\det(\Sigma_{B \cdot A})} \\
&= \frac{1}{2} \log \det \left( I_d + \Sigma_{\mathcal{M}(X)} \Sigma_B^{-1} \right) \\
&= \text{LogDet}(\mathcal{M}(X), B).
\end{aligned}$$

Therefore, Efficient-PAC implements a budget  $\beta$  that upper-bounds the exact Gaussian  $\text{LogDet}(\mathcal{M}(X), B)$ , with conser-



vativeness decomposing into the Hadamard step and the  $\log(1+x) \leq x$  linearization.

**Remark 1.** All our comparisons of Auto-PAC and SR-PAC that rely on the conservativeness of  $\text{LogDet}(\mathcal{M}(X), B)$  carry over verbatim for Efficient-PAC because Efficient-PAC implements  $\text{LogDet}(\mathcal{M}(X), B) \leq \beta$ , where the inequality is in general non-attainable. Thus, conservativeness-related results for  $\text{LogDet}(\mathcal{M}(X), B)$  remain valid a fortiori for the  $\beta$  implemented by Efficient-PAC.

**Remark 2.** Given any privacy budget, the upper bound implemented by Efficient-PAC induces more conservativeness than directly implementing  $\text{LogDet}(\mathcal{M}(X), B)$ . However, there is no universal ordering between the true mutual informations  $\text{MI}(X; \mathcal{M}(X) + B_{\text{Auto}})$  and  $\text{MI}(X; \mathcal{M}(X) + B_{\text{Eff}})$ , where  $B_{\text{Auto}}$  and  $B_{\text{Eff}}$  are the Gaussian noise determined by Auto-PAC and Efficient-PAC for the same privacy budget. This is because the Gaussianity gaps (explicitly formulated by (3)) of Auto-PAC and Efficient-PAC can be in general different magnitudes.

## B Membership Inference Attack

We first recall the standard definition of membership inference attacks formalized to match PAC Privacy [43, 46].

**Definition 10** (Membership Inference Attack [43, 46]). Given a finite data pool  $\mathcal{U} = \{u_1, u_2, \dots, u_N\}$  and some processing mechanism  $\mathcal{M}$ ,  $X$  is an  $n$ -subset of  $\mathcal{U}$  randomly selected. An informed adversary is asked to return an  $n$ -subset  $\hat{X}$  as the membership estimation of  $X$  after observing  $\mathcal{M}(X)$ . We say  $\mathcal{M}$  is resistant to  $(1 - \delta_i)$  individual membership inference for the  $i$ -th datapoint  $u_i$ , if for an arbitrary adversary,

$$\Pr_{X \leftarrow \mathcal{U}, \hat{X} \leftarrow \mathcal{M}(X)} (\mathbf{1}_{u_i \in X} = \mathbf{1}_{u_i \in \hat{X}}) \leq 1 - \delta_i$$

Here,  $\mathbf{1}_{u_i \in X}$  ( $\mathbf{1}_{u_i \in \hat{X}}$ ) is an indicator which equals 1 if  $u_i$  is in  $X$  ( $\hat{X}$ ).

Building on this attack model, we now introduce the corresponding R-PAC membership privacy notion:

**Definition 11** (R-PAC Membership Privacy). For a data processing mechanism  $\mathcal{M}$ , given some measure  $\rho$  and a data set  $\mathcal{U} = (u_1, u_2, \dots, u_N)$ , we say  $\mathcal{M}$  satisfies  $(R_f^\delta, \rho, \mathcal{U}, \mathcal{D})$ -R-PAC Membership Privacy if it is  $(\delta, \rho, \mathcal{U}, \mathcal{D})$  PAC Membership private and:

$$R_f^\delta \equiv \text{IntP}_f(\mathcal{D}) - D_f(\mathbf{1}_\delta \| \mathbf{1}_{\delta^o}) \quad (14)$$

is the posterior disadvantage, where:

- $\text{IntP}_f(\mathcal{D}) = -D_f(\mathcal{D} \| \mathcal{U})$  is the intrinsic membership privacy of the sampling distribution  $\mathcal{D}$  relative to the uniform distribution over  $\mathcal{U}$ ,

- $\mathbf{1}_\delta$  and  $\mathbf{1}_{\delta^o}$  denote the posterior and prior inference outcomes, respectively (thought of as binary success/failure indicators; equivalently, Bernoulli distributions with success parameters  $1 - \delta$  and  $1 - \delta^o$ ),
- $\delta_p^o = \inf_{\tilde{\mathbf{1}}_{\mathcal{U}}} \Pr_{X \sim \mathcal{D}}[\rho(\tilde{\mathbf{1}}_{\mathcal{U}}, \mathbf{1}_{\mathcal{U}}) \neq 1]$  is the optimal prior error level (so the optimal prior success is  $1 - \delta_p^o$ ).

$R_f^\delta$  quantifies the R-PAC membership privacy that persists after adversarial inference. The total intrinsic membership privacy is decomposed as:

$$\text{IntP}_f(\mathcal{D}) = R_f^\delta + \Delta_f^\delta, \quad \text{where } \Delta_f^\delta = D_f(\mathbf{1}_\delta \| \mathbf{1}_{\delta^o}) \quad (15)$$

is the PAC Membership Privacy loss, providing a complete accounting of membership privacy risk.

**KL case and Markov-chain justification.** When  $D_f$  is the KL divergence, write  $Y = \mathcal{M}(X)$  and let  $U_i \in \{0, 1\}$  be the membership indicator for an individual  $i$ , and  $J \in \{1, \dots, N\}$  the one-hot index with distribution  $\mathcal{D}$  (so  $\mathbf{1}_{\mathcal{U}}$  is the one-hot representation of  $J$ ). Then

$$\Delta_f^\delta = \text{KL}(\mathbf{1}_\delta \| \mathbf{1}_{\delta^o}) \leq I(U_i; Y) \leq I(J; Y),$$

where the first inequality is the Bernoulli-KL information-risk bound for membership, and the second follows because  $U_i$  is a deterministic function of  $J$  and  $U_i \rightarrow J \rightarrow X \rightarrow Y$  is a Markov chain (data processing). Moreover,

$$\text{IntP}_{\text{KL}}(\mathcal{D}) = \mathcal{H}(\mathcal{D}) - \log |\mathcal{U}|.$$

Combining these,

$$\begin{aligned} R_f^\delta &= \text{IntP}_{\text{KL}}(\mathcal{D}) - \Delta_f^\delta \geq \mathcal{H}(J|Y) - \log |\mathcal{U}| \\ &= \mathcal{H}(\mathbf{1}_{\mathcal{U}} | \mathcal{M}(X)) - V, \end{aligned}$$

with  $V \equiv \log |\mathcal{U}|$  independent of both  $\mathcal{D}$  and  $\mathcal{M}$ . This shows that the residual term lower-bounds the conditional uncertainty of the one-hot membership indicator given the mechanism output, up to a constant that depends only on the universe size.

## C Discussion: PAC/R-PAC Privacy vs. Differential Privacy

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

Unlike DP, which protects the privacy of individual records through worst-case probabilistic indistinguishability guarantees that must hold for any neighboring datasets, PAC Privacy aims to protect against arbitrary adversarial inference tasks under instance-based conditions. While DP focuses on ensuring that an adversary cannot distinguish whether any specific

individual participated in a dataset (treating this as a binary hypothesis testing problem), PAC Privacy provides a more general framework that quantifies the reconstruction hardness for any sensitive information that an adversary might seek to infer. Specifically, PAC Privacy characterizes the impossibility of an adversary successfully recovering sensitive data  $\tilde{X}$  such that  $\rho(X, \tilde{X}) = 1$  under any user-defined criterion  $\rho$ , given observation of the mechanism output  $\mathcal{M}(X)$ . This encompasses not only individual membership inference (a special case), but also broader privacy concerns such as data reconstruction within specified error bounds, identification of multiple participants, or recovery of sensitive attributes. Crucially, PAC Privacy operates under distributional assumptions about the data generation process  $D$ , enabling instance-based analysis that can potentially require less noise than worst-case DP guarantees, while providing semantic security interpretation through concrete bounds on adversarial success probabilities for any reconstruction objective.

R-PAC and PAC Privacy are two sides of the same coin: PAC quantifies leakage (e.g., via  $\Delta_f^\delta$ ), while R-PAC quantifies the remaining privacy (e.g., via  $R_f^\delta$ ), linked exactly by  $\text{IntP}_f(\mathcal{D}) = R_f^\delta + \Delta_f^\delta$  (equation (9)). PAC Privacy (as a *definition*) are adversary-agnostic and computation-unbounded: the stated guarantees bound an attacker’s advantage for *any* inference strategy, without restricting computational power, and they are formulated *without* explicitly using mutual information (MI), Fisher information, or other information-theoretic metrics.

However, the automatic privatization procedures (e.g., Auto-PAC and Efficient-PAC) proposed to realize PAC Privacy certify the privacy guarantee via an MI budget  $\beta$ , enforcing Gaussian surrogate bound  $\text{LogDet}(\mathcal{M}(X), B) \leq \beta$  as a sufficient condition. Consequently, the delivered mechanisms inherit MI-based properties and caveats (e.g., data processing, distributional/average-case nature, standard composition scaling, and lack of worst-case indistinguishability unless additional constraints are imposed), even though the abstract PAC Privacy notion itself does not rely on MI. Our SR-PAC follows the MI principle but implements an MI bound that is tighter than the Gaussian surrogate bound  $\text{LogDet}(\mathcal{M}(X), B)$ .

In the next section, we discuss the difference between MI privacy and DP. Unless stated otherwise, we focus on PAC Privacy in this discussion for simplicity.

## C.1 Mutual Information Privacy vs. Differential Privacy

**What each notion protects.** Let  $\mathcal{M} : \mathcal{X} \mapsto \mathcal{Y}$  be a randomized mechanism. DP, independent of input distribution, protects *worst-case, per-individual indistinguishability* by ensuring a uniform bound  $\ell(x, y) \leq \epsilon$  almost surely (up to  $\delta$  in the  $(\epsilon, \delta)$  case), where  $\ell(x, y) = \log \frac{P_{\mathcal{M}(Y)}(x|y)}{P_X(x)}$  is the privacy-loss random variable. However, DP does not, in general, en-

sure that the *average* leakage  $\text{MI}(X; Y)$  is small—indeed,  $\text{MI}(X; Y)$  can scale with the dataset size unless  $\epsilon$  shrinks appropriately. In contrast, MI-based privacy constrains the *average information* leaked from inputs to outputs under a specific input distribution  $\mathcal{D} \in \Delta(\mathcal{X})$ . MI controls average leakage:  $\text{MI}(X; Y) = \mathbb{E}_{P_{XY}}[\ell(X, Y)] \leq \beta$  upper-bounds the expected log-likelihood gain of an optimal Bayesian adversary. However, MI *does not* by itself bound the worst-case leakage  $L \triangleq \text{ess sup } \ell$ ; in particular,  $\text{MI}(X; Y) \leq \beta$  is compatible with  $L = \infty$  (rare but arbitrarily large disclosures).

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

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

**DP perspective on the name-and-shame example.** Consider the per-record "name-and-shame" mechanism  $M$  that, independently for each index  $i$ , outputs  $(i, x_i)$  with probability  $p$  and  $\perp$  otherwise. Let  $x, x'$  be neighboring databases that differ only in record  $i$  with  $x_i \neq x'_i$ , and define the event  $E = (i, x_i)$ . Then  $\Pr[M(x) \in E] = p$ ,  $\Pr[M(x') \in E] = 0$ . The  $(\epsilon, \tilde{\delta})$ -DP inequality for  $E$  reads  $p \leq e^\epsilon \cdot 0 + \tilde{\delta} = \tilde{\delta}$ , hence any  $(\epsilon, \tilde{\delta})$  satisfied by  $M$  must obey  $\tilde{\delta} \geq p$ . In particular, with the

standard regime  $\bar{\delta} \ll 1/n$  (negligible failure probability), such a mechanism is *not* DP for any finite  $\epsilon$ ; conversely, allowing  $\bar{\delta} \geq p$  makes the guarantee vacuous on the  $p$ -fraction of runs that reveal  $(i, x_i)$  exactly.

## C.2 Fair Comparison Under MIA

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

DP can be understood through the lens of membership inference success rates. Consider the membership inference scenario from Definition 10, where we have a dataset of size  $n = \frac{N}{2}$  (i.e., each individual data record has a 50% probability of being included in the selected subset  $X$ ). If a mechanism  $\mathcal{M}$  is  $(\epsilon, \bar{\delta})$ -DP, then by [25, 28], an adversary's ability to successfully infer whether a specific individual record  $i$  is included in the dataset (i.e., posterior success rate  $p_o = 1 - \bar{\delta}_i$ ) is fundamentally limited:

$$p_o \leq 1 - \frac{1 - \bar{\delta}}{1 + e^\epsilon}. \quad (16)$$

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

In addition, there is a relationship between the posterior success rate  $p_o$  and the mutual information [43] (derived from (1)):

$$p_o \log \frac{p_o}{\bar{p}} + (1 - p_o) \log \frac{1 - p_o}{1 - \bar{p}} \leq \text{MI}(X; \mathcal{M}(X)), \quad (17)$$

where  $\bar{p}$  is the optimal prior success rate, which is  $\max(r, 1 - r)$  with  $r$  as the subsampling rate that selects the dataset from a data pool. Thus, given a privacy budget  $\text{MI}(X; \mathcal{M}(X)) = \beta$  and a prior success rate  $\bar{p}$ , we can calculate the posterior success level  $p_o$  and, by (16), pin down  $\epsilon$  for a chosen  $\bar{\delta}$  so that DP has an "equivalent" budget to PAC. The corresponding R-PAC budget is  $\mathcal{H}(X) - \beta$ .

For per-individual membership, the relevant secret is the inclusion bit  $U_i \in \{0, 1\}$  for person  $i$  and the mechanism output is  $Y = \mathcal{M}(X)$ . Since  $U_i \rightarrow X \rightarrow Y$  forms a Markov chain, the data-processing inequality gives

$$\text{MI}(U_i; Y) \leq \text{MI}(X; Y) = \beta.$$

The Bernoulli–KL inequality used above applies equally with  $\text{MI}(U_i; Y)$  on the right-hand side; replacing it by  $\text{MI}(X; Y)$  is therefore conservative and still yields a valid upper bound on the Bayes-optimal membership posterior success  $p_o$ . This

validates using  $\text{MI}(X; Y)$  to compute  $p_o(\beta, \bar{p})$  for MIA and then selecting  $(\epsilon, \bar{\delta})$  so that (16) enforces the same  $p_o$  for a fair, like-for-like comparison between DP and PAC/R-PAC.

## C.3 Noise Magnitude

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

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

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

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

$$\left(\sum_{j=1}^p \sqrt{\lambda_j}\right)^2 \leq p \sum_{j=1}^p \lambda_j \leq p^2 R, \quad R \equiv \max_j \lambda_j,$$

hence  $V_{\text{PAC}}(\beta) = O(p^2 R/\beta)$  in the worst case (and improves to  $O(pR/\beta)$  if  $\sum_j \lambda_j = O(R)$ ). (In practice, Algorithm 1 uses *estimated eigenvalues* and a *stabilization*  $10cv/\beta$ , yielding *total noise magnitude*  $(\sum_j \sqrt{\hat{\lambda}_j + 10cv/\beta})^2/(2v)$ , which is a *conservative upper envelope of the ideal log-det calibration*.) Because differential privacy (DP) must mask *worst-case* changes in all  $d$  coordinates, the required noise for  $d$ -dimensional outputs typically grows like  $\sqrt{d}$  (e.g.,  $O(\sqrt{d}/n)$  for mean queries with dataset size  $n$ )—the classic "curse of dimensionality." Thus, when the data are effectively low-rank ( $p \ll d$ ), Auto-PAC already mitigates this dimensional blow-up.

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

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

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

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

- *Anisotropic allocation:* The Stackelberg-optimal covariance is provably anisotropic; variance is shifted toward directions with high leakage and away from benign ones, improving utility without violating the MI budget.
- *Zero-noise subspaces:* Under a mild separation of directional sensitivities, there exists a threshold  $\beta_{\text{lab}}$  such that for all  $\beta \leq \beta_{\text{lab}}$  SR-PAC injects *no* noise on an  $s$ -dimensional task-critical subspace (e.g., the  $k-1$  label directions in classification), reducing the order from  $O(p)$  to  $O(p-s)$  in those regimes.

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

$$V_{\text{SR}}(\beta) = V_{\text{PAC}}(\beta) - \Delta(\beta), \quad 0 \leq \Delta(\beta) \leq V_{\text{PAC}}(\beta),$$

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

## D Technical Constructions of Reference Distributions for Intrinsic Privacy

**Purpose and scope.** Intrinsic privacy is defined as

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

where  $\mathcal{R}$  is a *reference distribution* that plays the role of an a priori baseline and  $D_f$  is an  $f$ -divergence (KL in our evaluations). To make  $\text{IntP}_f$  well-defined and mechanism-independent, one must choose  $\mathcal{R}$  so that (i)  $\text{supp}(\mathcal{D}) \subseteq \text{supp}(\mathcal{R})$  and (ii)  $D_f(\mathcal{D}||\mathcal{R}) < \infty$ . This appendix gives three canonical constructions of  $\mathcal{R}$  together with conditions that guarantee finiteness, and brief practical advice on when to use each choice.

**Standing notation.** We write  $X \sim \mathcal{D}$  for the data distribution on  $\mathbb{R}^d$ . A reference distribution  $\mathcal{R}$  has density  $r(\cdot)$  w.r.t. Lebesgue measure (whenever it exists). For KL,  $D_{\text{KL}}(\mathcal{D}||\mathcal{R}) = \mathbb{E}_{\mathcal{D}}[\ln \frac{d\mathcal{D}}{d\mathcal{R}}(X)]$  and  $H(\mathcal{R})$  denotes the (differential) entropy of  $\mathcal{R}$  (log base as in the main text).

**Proposition 8** (Finiteness criteria for KL). *If  $\mathcal{D} \ll \mathcal{R}$  and  $\mathbb{E}_{\mathcal{D}}[|\ln r(X)|] < \infty$ , then  $D_{\text{KL}}(\mathcal{D}||\mathcal{R}) < \infty$ . Consequently, any construction of  $\mathcal{R}$  that ensures full support on  $\mathbb{R}^d$  and mild tail control on  $r$  suffices for finiteness of  $\text{IntP}_{\text{KL}}$ .*

*Proof.* Since  $\mathcal{R}$  has Lebesgue density  $r$  and  $\mathcal{D} \ll \mathcal{R}$ , we also have  $\mathcal{D} \ll \text{Lebesgue}$ ; let  $p$  denote the Lebesgue density of  $\mathcal{D}$ . By the chain rule for Radon–Nikodym derivatives,

$$\frac{d\mathcal{D}}{d\mathcal{R}}(x) = \frac{d\mathcal{D}/dx}{d\mathcal{R}/dx}(x) = \frac{p(x)}{r(x)} \quad \text{a.e.}$$

Hence

$$\begin{aligned} D_{\text{KL}}(\mathcal{D}||\mathcal{R}) &= \int p(x) \ln \frac{p(x)}{r(x)} dx \\ &= \underbrace{\int p(x) \ln p(x) dx}_{-H(\mathcal{D})} - \underbrace{\int p(x) \ln r(x) dx}_{\mathbb{E}_{\mathcal{D}}[\ln r(X)]}. \end{aligned}$$

By assumption,  $H(\mathcal{D}) > -\infty$  and  $\mathbb{E}_{\mathcal{D}}[|\ln r(X)|] < \infty$ , so both terms on the right-hand side are finite (the first from below, the second in absolute value), and their difference is finite. Therefore  $D_{\text{KL}}(\mathcal{D}||\mathcal{R}) < \infty$ .  $\square$

In the KL case, our residual privacy lower bound involves a constant offset  $V = H(\mathcal{R})$  (independent of both  $\mathcal{D}$  and the mechanism), so we also highlight when  $H(\mathcal{R}) < \infty$ .

### (a) Maximum-entropy Gaussian

The maximum-entropy Gaussian is defined as

$$\mathcal{R} = \mathcal{N}(\mu, \Sigma), \quad \mu = \mathbb{E}_{\mathcal{D}}[X], \quad \Sigma = \text{Cov}_{\mathcal{D}}(X).$$

The density function is

$$r(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right),$$

with the support  $\text{supp}(\mathcal{R}) = \mathbb{R}^d$ . The corresponding entropy is

$$H(\mathcal{R}) = \frac{1}{2} \ln((2\pi e)^d \det \Sigma) < \infty.$$

If  $\mathcal{D}$  is absolutely continuous and  $\mathbb{E}_{\mathcal{D}}[\|X\|^2] < \infty$ , then  $D_{\text{KL}}(\mathcal{D}||\mathcal{R}) < \infty$ .

It is a natural default when second moments exist; full support guarantees  $\text{supp}(\mathcal{D}) \subseteq \text{supp}(\mathcal{R})$  automatically. In practice, ensure  $\Sigma \succ 0$  via standard shrinkage if needed.

### (b) Smooth pull-back of the unit-cube uniform

The smooth pull-back construction is defined as follows: let  $U \sim \text{Unif}((0,1)^d)$  and choose a  $C^1$  bijection

$$T : (0,1)^d \rightarrow \mathbb{R}^d, \quad \det J_T(u) > 0.$$



The reference is the push-forward  $\mathcal{R} = T_{\#}U$  with density

$$r(x) = |\det J_{T^{-1}}(x)|,$$

and support  $\text{supp}(\mathcal{R}) = \mathbb{R}^d$ . The corresponding entropy is

$$H(\mathcal{R}) = \mathbb{E}_U [\ln |\det J_T(U)|] < \infty$$

whenever  $\ln |\det J_T|$  is integrable on  $(0, 1)^d$ . If  $\mathcal{D} \ll \mathcal{R}$  and  $\mathbb{E}_{\mathcal{D}}[|\ln r(X)|] < \infty$ , then  $D_{\text{KL}}(\mathcal{D} \parallel \mathcal{R}) < \infty$ .

It is useful when one wishes to encode geometry or tail behavior via the map  $T$  while retaining full support and finite  $H(\mathcal{R})$  through an integrability check on  $\ln |\det J_T|$ .

### (c) Truncated uniform on a bounded set

The truncated uniform is defined as follows: let  $B \subset \mathbb{R}^d$  be compact with  $\text{supp}(\mathcal{D}) \subseteq B$ , and set

$$\mathcal{R} = \text{Unif}(B), \quad r(x) = \begin{cases} 1/\text{vol}(B), & x \in B, \\ 0, & x \notin B. \end{cases}$$

The support is  $\text{supp}(\mathcal{R}) = B$ . The corresponding entropy is

$$H(\mathcal{R}) = \ln(\text{vol}(B)) < \infty.$$

Moreover,

$$D_{\text{KL}}(\mathcal{D} \parallel \mathcal{R}) = -H(\mathcal{D}) + \ln(\text{vol}(B)),$$

so finiteness requires  $H(\mathcal{D}) < \infty$ .

It is appropriate only when the domain is naturally bounded and the data distribution has finite entropy; otherwise, one should prefer the Gaussian or pull-back constructions.

## E More on Non-Gaussianity Correction

In Section 3.2, we propose two approaches to approximate the Gaussianity gap  $\text{Gap}_a$ , which are certified replacements of  $D_Z$  to find a tighter mutual information after Auto-PAC privatization. Theorem 3 uses Donsker–Varadhan (DV) representation  $D_Z = \sup_f \{\mathbb{E}_{P_{\mathcal{M},B}} f - \log \mathbb{E}_{\tilde{Q}_{\mathcal{M}}} e^f\}$ , so that any value of the DV objective at a trained critic  $f_{\psi}$  is a valid lower bound on  $D_Z$ . Under a mild transport condition for  $P_{\mathcal{M},B}$  and  $\tilde{Q}_{\mathcal{M}}$ , Theorem 4 use the sliced Wasserstein distance (SWD) as the estimation  $\hat{D}_Z$ , which is unbiased in the minibatch limit. In addition, the estimation achieves a certified  $0 \leq \hat{D}_Z \leq D_Z$ .

Consequently, our improved mutual information estimate

$$\text{IMI}(\hat{D}_Z) = \text{LogDet}(\mathcal{M}(X), B) - \hat{D}_Z$$

is a provable upper bound on  $\text{MI}(X; Z)$  whenever  $\hat{D}_Z$  is one of the certified corrections above.

Both approaches admit short, minibatch estimators:

---

### Algorithm 4 DV Gap Correction (minibatch lower bound on $D_Z$ )

---

**Require:** Oracle for i.i.d. samples  $Z \sim P_{\mathcal{M},B}$ ; function class  $\mathcal{F} = \{f_{\phi}\}$ ; steps  $T$ ; batch size  $m$ ; step size  $\eta$ ; confidence penalty  $c_{m,\delta}$

- 1: Draw an initial batch  $\{Z_i\}_{i=1}^{m_0} \sim P_{\mathcal{M},B}$  and estimate  $\hat{\mu}_Z, \hat{\Sigma}_Z$ ; define  $\tilde{Q}_{\mathcal{M}} := \mathcal{N}(\hat{\mu}_Z, \hat{\Sigma}_Z)$
  - 2: Initialize  $\phi$
  - 3: **for**  $t = 1, \dots, T$  **do**
  - 4:   Sample  $\{Z_j^{(P)}\}_{j=1}^m \sim P_Z$
  - 5:   Sample  $\{Z_j^{(Q)}\}_{j=1}^m \sim \tilde{Q}_{\mathcal{M}}$
  - 6:    $\hat{b} \leftarrow \frac{1}{m} \sum_{j=1}^m f_{\phi}(Z_j^{(P)}) - \log \left( \frac{1}{m} \sum_{j=1}^m e^{f_{\phi}(Z_j^{(Q)})} \right)$
  - 7:    $\phi \leftarrow \phi + \eta \nabla_{\phi} \hat{b}$
  - 8: **end for**
  - 9: Evaluate  $\hat{b}_{\text{val}}$  on held-out minibatches; set  $\underline{D}_Z \leftarrow \max\{\hat{b}_{\text{val}} - c_{m,\delta}, 0\}$
  - 10: **Return**  $\underline{D}_Z$
- 

- **DV Correction.** Train a critic  $f_{\phi}$  by maximizing

$$\hat{\mathcal{J}} = \frac{1}{m} \sum_{i=1}^m f_{\phi}(Z_i) - \log \left( \frac{1}{m} \sum_{i=1}^m e^{f_{\phi}(\tilde{Z}_i)} \right),$$

where  $Z_i \sim P_Z$  and  $\tilde{Z}_i \sim \tilde{Q}_{\mathcal{M}}$ . After  $T$  steps, set  $\hat{D}_Z \leftarrow \hat{\mathcal{J}}(f_{\phi})$ . Algorithm 4 shows an example.

- **SWD Correction.** Sample  $K$  directions  $v_k \sim \text{Unif}(\mathbb{S}^{d-1})$ , project both batches, sort each projection, and average 1D squared distances:

$$\widehat{\text{SW}}_2^2 = \frac{1}{K} \sum_{k=1}^K \frac{1}{m} \sum_{j=1}^m (\langle v_k, Z \rangle_{(j)} - \langle v_k, \tilde{Z} \rangle_{(j)})^2.$$

Convert  $\widehat{\text{SW}}_2^2$  to  $\hat{D}_Z$  using the calibration stated in Theorem 4. Algorithm 5 gives an example.

Each iteration uses a single minibatch pass and either a small critic update (DV) or  $K$  sorts of length  $m$  (SWD); no back-propagation through  $\mathcal{M}$  and no nested inner loops.

## F PAC Generalization Bound for the Follower

Fix a perturbation rule  $Q \in \Gamma$ . The Follower’s objective is

$$\pi^*(Q) \in \arg \min_{\pi \in \Pi} W(Q, \pi),$$

where

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

**Algorithm 5** Sliced Wasserstein Gap Correction (training-free lower bound on  $D_Z$ )

**Require:** Oracle for i.i.d. samples  $Z \sim P_{\mathcal{M},B}$ ; number of projections  $M$ ; samples per slice  $n$ ; confidence penalty  $\xi_{n,\delta}$

- 1: Draw an initial batch  $\{Z_i\}_{i=1}^{n_0} \sim P_{\mathcal{M},B}$  and estimate  $\hat{\mu}_Z, \hat{\Sigma}_Z$ ;  
set  $W \leftarrow \hat{\Sigma}_Z^{-1/2}$
- 2: **for**  $m = 1, \dots, M$  **do**
- 3: Draw  $\theta_m$  uniformly on  $\mathbb{S}^{d-1}$
- 4: Draw  $n$  fresh samples  $Z_i \sim P_Z$  and set  $u_i = \theta_m^\top W(Z_i - \hat{\mu}_Z)$
- 5: Draw  $n$  i.i.d. samples  $s_i \sim \mathcal{N}(0, 1)$
- 6: Sort  $u_{(1)} \leq \dots \leq u_{(n)}$  and  $s_{(1)} \leq \dots \leq s_{(n)}$ ; set  $w_m^2 \leftarrow \frac{1}{n} \sum_{i=1}^n (u_{(i)} - s_{(i)})^2$
- 7: **end for**
- 8:  $\widehat{SW}_2^2 \leftarrow \frac{1}{M} \sum_{m=1}^M w_m^2$
- 9:  $\underline{D}_Z \leftarrow \max\{\frac{1}{2}\widehat{SW}_2^2 - \xi_{n,\delta}, 0\}$
- 10: **Return**  $\underline{D}_Z$

Given  $m$  i.i.d. samples  $(X_i, B_i, Y_i)_{i=1}^m$  with  $X_i \sim \mathcal{D}$ ,  $B_i \sim Q$ , and  $Y_i = M(X_i) + B_i$ , define the empirical risk

$$\widehat{W}(Q, \pi) = \frac{1}{m} \sum_{i=1}^m [-\log \pi(X_i | Y_i)], \quad \hat{\pi} \in \arg \min_{\pi \in \Pi} \widehat{W}(Q, \pi).$$

Let  $G_\Pi \equiv \{g_\pi(x, y) = -\log \pi(x | y) : \pi \in \Pi\}$  and denote by  $\mathcal{R}_m(G_\Pi)$  the empirical Rademacher complexity of  $G_\Pi$  on  $m$  samples.

**Assumption (bounded log-likelihood).** There exists  $B > 0$  such that for all  $\pi \in \Pi$  and all  $(x, y)$  in the support,  $-\log \pi(x | y) \in [0, B]$ . (When densities are unbounded, this is enforced by standard truncation or by lower-bounding the decoder's variance / softmax temperature over a bounded input domain.)

**Lemma 1** (Decoder PAC generalization). *Under the boundedness assumption, for any  $\delta \in (0, 1)$ , with probability at least  $1 - 2\delta$  over the draw of the  $m$  samples,*

$$\left| \inf_{\pi \in \Pi} W(Q, \pi) - \widehat{W}(Q, \hat{\pi}) \right| \leq \underbrace{4 \mathcal{R}_m(G_\Pi)}_{\text{capacity}} + \underbrace{2B \sqrt{\frac{2 \log(1/\delta)}{m}}}_{\text{concentration}} \equiv \epsilon_{m,\delta}.$$

*Proof.* Let  $G_\Pi = \{g_\pi(x, y) = -\log \pi(x | y) : \pi \in \Pi\}$  with  $g_\pi \in [0, B]$  by assumption, and let

$$\widehat{\mathcal{R}}_m(G_\Pi) \equiv \mathbb{E}_\sigma \left[ \sup_{g \in G_\Pi} \frac{1}{m} \sum_{i=1}^m \sigma_i g(X_i, Y_i) \right]$$

be the (empirical) Rademacher complexity on the sample  $(X_i, Y_i)_{i=1}^m$ , where  $\sigma_i \in \{\pm 1\}$  are i.i.d. Rademacher variables. By standard symmetrization and McDiarmid's inequality

(bounded differences  $B/m$ ), with probability at least  $1 - \delta$ ,

$$\sup_{\pi \in \Pi} |W(Q, \pi) - \widehat{W}(Q, \pi)| \leq 2 \widehat{\mathcal{R}}_m(G_\Pi) + B \sqrt{\frac{2 \log(1/\delta)}{m}}. \quad (18)$$

On the same event, let  $\pi^* \in \arg \min_{\pi} W(Q, \pi)$  and  $\hat{\pi} \in \arg \min_{\pi} \widehat{W}(Q, \pi)$ . Then

$$\begin{aligned} \inf_{\pi} W(Q, \pi) - \widehat{W}(Q, \hat{\pi}) &= W(Q, \pi^*) - \widehat{W}(Q, \hat{\pi}) \\ &\geq -\sup_{\pi} |W - \widehat{W}| - \sup_{\pi} |W - \widehat{W}| \\ &\geq -2\Delta, \end{aligned}$$

where  $\Delta = 2 \widehat{\mathcal{R}}_m(G_\Pi) + B \sqrt{2 \log(1/\delta)/m}$  is the right-hand side of (18). Thus

$$\begin{aligned} \left| \inf_{\pi \in \Pi} W(Q, \pi) - \widehat{W}(Q, \hat{\pi}) \right| &\leq 2\Delta \\ &= 4 \widehat{\mathcal{R}}_m(G_\Pi) + 2B \sqrt{\frac{2 \log(1/\delta)}{m}}. \end{aligned}$$

Finally, applying a standard concentration step to replace  $\widehat{\mathcal{R}}_m(G_\Pi)$  by its expectation  $\mathcal{R}_m(G_\Pi)$  incurs an additional failure probability  $\delta$ ; a union bound yields the stated result with probability at least  $1 - 2\delta$ .  $\square$

**Corollary 4** (Finite-sample feasibility check for the Leader). *Let  $\hat{\beta}$  be the residual-PAC budget in the Leader's constraint  $\inf_{\pi \in \Pi} W(Q, \pi) \geq \hat{\beta}$ . If the batch cross-entropy  $H_c = \widehat{W}(Q, \hat{\pi})$  (the quantity computed in Algorithm 3) satisfies*

$$H_c \geq \hat{\beta} + \epsilon_{m,\delta},$$

*then, with probability at least  $1 - 2\delta$ ,  $\inf_{\pi \in \Pi} W(Q, \pi) \geq \hat{\beta}$ .*

**PAC-adjusted penalty.** Define the PAC-adjusted threshold

$$\hat{\beta}_{\text{PAC}} \equiv \hat{\beta} + \epsilon_{m,\delta}.$$

A convenient implementation is to use  $\hat{\beta}_{\text{PAC}}$  in place of  $\hat{\beta}$  inside the Leader's penalty; i.e., set

$$\text{penalty} = \sigma (H_c - \hat{\beta}_{\text{PAC}})_+^2,$$

where  $\sigma$  is the penalty weight and  $(\cdot)_+ = \max\{\cdot, 0\}$ . Algorithm 6 shows the SR-PAC with the PAC-adjusted penalty. By Corollary 4, any iterate with zero penalty (or sufficiently small penalty when smooth proxies are used) satisfies the population constraint with probability at least  $1 - 2\delta$  at sample size  $m$ .

---

**Algorithm 6** Monte Carlo SR-PAC (with PAC-adjusted Penalty)

---

**Require:** Privacy budget  $\hat{\beta}$ , parametrized decoder family  $\Pi_\phi$ , perturbation rule family  $\Gamma_\lambda$ , utility loss  $\mathcal{K}(\cdot)$ , learning rates  $\eta_\phi, \eta_\lambda$ , penalty weight  $\sigma$ , iterations  $T_\lambda, T_\phi$ , batch size  $m$

```

1: Initialize parameters  $\lambda, \phi \sim \text{init}()$ 
2: for  $t = 1, \dots, T_\lambda$  do
3:   if  $t \bmod T_\phi = 0$  then
4:     Update Decoder:
5:     for  $i = 1, \dots, T_\phi$  do
6:       Sample  $\{(x_j, b_j, y_j)\}_{j=1}^m$  where  $x_j \sim \mathcal{D}, b_j \sim Q_\lambda,$ 
7:          $y_j = \mathcal{M}(x_j) + b_j$ 
8:        $\hat{W} = \frac{1}{m} \sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$ 
9:        $\phi \leftarrow \phi - \eta_\phi \nabla_\phi \hat{W}$ 
10:    end for
11:  end if
12:  Update Perturbation Rule:
13:  Sample  $\{(x_j, b_j, y_j)\}_{j=1}^m$  where  $x_j \sim \mathcal{D}, b_j \sim Q_\lambda, y_j =$ 
14:     $\mathcal{M}(x_j) + b_j$ 
15:   $H_c = \frac{1}{m} \sum_{j=1}^m [-\log \pi_\phi(x_j|y_j)]$ 
16:   $\mathcal{L}_\lambda = \frac{1}{m} \sum_{j=1}^m \mathcal{K}(b_j) + \sigma(H_c - (\hat{\beta} + \epsilon_{m,\delta}))_+^2$ 
17:   $\lambda \leftarrow \lambda - \eta_\lambda \nabla_\lambda \mathcal{L}_\lambda$ 
18: end for
19: return Optimal parameters  $(\lambda^*, \phi^*)$ 

```

---

**Sample complexity (reading  $\epsilon_{m,\delta}$ ).** If  $\mathcal{R}_m(G_\Pi) \leq C/\sqrt{m}$ , then any

$$m \geq \left( \frac{4C}{\eta} + \frac{2B\sqrt{2\log(1/\delta)}}{\eta} \right)^2$$

ensures  $\epsilon_{m,\delta} \leq \eta$ , so enforcing  $H_c \geq \hat{\beta} + \eta$  certifies feasibility with probability  $\geq 1 - 2\delta$ . By Corollary 4, this appendix-only variant certifies feasibility with probability at least  $1 - 2\delta$  at sample size  $m$ .

## G Proof of Theorem 3

Let  $Z = \mathcal{M}(X) + B$  with deterministic  $\mathcal{M}$  and  $B \sim \mathcal{N}(0, \Sigma_B)$  where  $\Sigma_B \succ 0$ . Write  $P_{\mathcal{M},B}$  for the law of  $Z$ . Let the Gaussian surrogate be  $\tilde{Q}_\mathcal{M} = \mathcal{N}(\mu_Z, \Sigma_Z)$ . For any measurable  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  with  $\mathbb{E}_{\tilde{Q}_\mathcal{M}}[e^f] < \infty$ , define the Donsker–Varadhan (DV) objective

$$\mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) = \mathbb{E}_{P_{\mathcal{M},B}}[f(Z)] - \log \mathbb{E}_{\tilde{Q}_\mathcal{M}}[e^{f(Z)}].$$

Let

$$\hat{\mathcal{D}}_Z(f) \equiv \mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}).$$

Next, we construct the finite-sample estimation of the DV objective. Given finite samples  $S_P$  from  $P_{\mathcal{M},B}$  and  $S_Q$  from

$\tilde{Q}_\mathcal{M}$ , let

$$\hat{\mathcal{J}}(f; S_P, S_Q) \equiv \frac{1}{|S_P|} \sum_{z \in S_P} f(z) - \log \left( \frac{1}{|S_Q|} \sum_{z \in S_Q} e^{f(z)} \right).$$

Fix a function class  $\mathcal{F} \subset \{f: \mathbb{R}^d \rightarrow \mathbb{R}\}$  with  $0 \in \mathcal{F}$ . Draw four independent splits  $S_P^{\text{tr}}, S_Q^{\text{tr}}, S_P^{\text{val}}, S_Q^{\text{val}}$  with sizes  $n_P^{\text{tr}}, n_Q^{\text{tr}}, n_P^{\text{val}}, n_Q^{\text{val}}$ , respectively, and fit

$$\hat{f}_{\text{tr}} \in \arg \max_{f \in \mathcal{F}} \hat{\mathcal{J}}(f; S_P^{\text{tr}}, S_Q^{\text{tr}}).$$

Assume that for some  $\Gamma_{\hat{\delta}} = \Gamma_{\hat{\delta}}(\mathcal{F}, n_P^{\text{val}}, n_Q^{\text{val}})$ ,

$$\Pr \left( \sup_{f \in \mathcal{F}} \left| \hat{\mathcal{J}}(f; S_P^{\text{val}}, S_Q^{\text{val}}) - \mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) \right| \leq \Gamma_{\hat{\delta}} \right) \geq 1 - \hat{\delta}. \quad (19)$$

Define the *finite-sample lower-confidence estimator*

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

**Proof of (i)** We apply the Gibbs variational principle. For any  $P \ll Q$  and measurable  $f$  with  $\mathbb{E}_Q[e^f] < \infty$ ,

$$\mathbb{E}_P[f] - \log \mathbb{E}_Q[e^f] \leq \text{D}_{\text{KL}}(P \| Q),$$

with equality at  $f^* = \log \frac{dP}{dQ} + c$  (any constant  $c$ ).

Taking the supremum over  $f$  yields

$$\sup_f \mathcal{J}(f; P, Q) = \text{D}_{\text{KL}}(P \| Q).$$

Apply with  $P = P_{\mathcal{M},B}$ ,  $Q = \tilde{Q}_\mathcal{M}$ . It is nonnegative because  $f \equiv 0$  is admissible and  $\mathcal{J}(0; P, Q) = 0$ .

**Proof of (ii)** This is immediate from Part (i):

$$\begin{aligned} \hat{\mathcal{D}}_Z(f) &= \mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) \leq \sup_g \mathcal{J}(g; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) \\ &= \text{D}_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_\mathcal{M}). \end{aligned}$$

**Proof of (iii)** Conditioning on the training splits  $(S_P^{\text{tr}}, S_Q^{\text{tr}})$ ,  $\hat{f}_{\text{tr}}$  (a measurable function of the training data) is independent of the validation splits  $(S_P^{\text{val}}, S_Q^{\text{val}})$ . On the event in (19), we have for all  $f \in \mathcal{F}$ ,

$$\mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) \geq \hat{\mathcal{J}}(f; S_P^{\text{val}}, S_Q^{\text{val}}) - \Gamma_{\hat{\delta}}.$$

Taking  $f = \hat{f}_{\text{tr}}$  and then the supremum over  $f$  on the left,

$$\text{D}_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_\mathcal{M}) = \sup_f \mathcal{J}(f; P_{\mathcal{M},B}, \tilde{Q}_\mathcal{M}) \geq \hat{\mathcal{J}}(\hat{f}_{\text{tr}}; S_P^{\text{val}}, S_Q^{\text{val}}) - \Gamma_{\hat{\delta}}.$$

Hence, on that event still,

$$0 \leq \left[ \hat{\mathcal{J}}(\hat{f}_{\text{tr}}; S_P^{\text{val}}, S_Q^{\text{val}}) - \Gamma_{\hat{\delta}} \right]_+ \leq \text{D}_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_\mathcal{M}),$$

which is the claim with probability at least  $1 - \hat{\delta}$ .

## H Proof of Theorem 4

Given the true (perturbed output) distribution  $P_{\mathcal{M},B}$  and the Gaussian surrogate distribution  $\tilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$  (with matched mean and covariance), define

$$\hat{D}_Z \equiv \frac{1}{2\lambda_{\max}(\Sigma_Z)} \text{SW}_2^2(P_{\mathcal{M},B}, \tilde{Q}_{\mathcal{M}}),$$

where  $\lambda_{\max}(\Sigma_Z)$  is the largest eigenvalue of  $\Sigma_Z$  and  $\text{SW}_2$  denotes the sliced 2-Wasserstein distance.

By definition of the Wasserstein metric,  $\text{SW}_2^2(P_{\mathcal{M},B}, \tilde{Q}_{\mathcal{M}}) \geq 0$ , and  $\lambda_{\max}(\Sigma_Z) > 0$  since  $\Sigma_Z$  is positive semidefinite and non-degenerate. Hence,

$$\hat{D}_Z \geq 0.$$

In addition, it is well known (see e.g., [3]) that the sliced Wasserstein distance provides a lower bound on the true 2-Wasserstein distance:

$$\text{SW}_2^2(P_Z, \tilde{Q}_{\mathcal{M}}) \leq W_2^2(P_Z, \tilde{Q}_{\mathcal{M}}).$$

Finally, Lemma 2 (shown below) implies

$$\frac{1}{2\lambda_{\max}(\Sigma_Z)} \text{SW}_2^2(P_{\mathcal{M},B}, \tilde{Q}_{\mathcal{M}}) \leq \text{D}_{\text{KL}}(P \| Q).$$

Therefore, we obtain

$$0 \leq \hat{D}_Z \leq \text{D}_Z$$

**Lemma 2.** Let  $Q = \mathcal{N}(\mu, \Sigma)$  with  $\Sigma \succ 0$  and let  $P$  be a probability measure on  $\mathbb{R}^d$  with  $P \ll Q$ . Then

$$W_2^2(P, Q) \leq 2\lambda_{\max}(\Sigma) \text{D}_{\text{KL}}(P \| Q).$$

*Proof.* Let  $T(x) = \Sigma^{-1/2}(x - \mu)$ . Then,  $T$  is invertible affine. In addition, let  $P' = T_{\#}P$  and  $\gamma = \mathcal{N}(0, I)$ . Thus, applying the change of variables yields

$$\text{D}_{\text{KL}}(P' \| \gamma) = \text{D}_{\text{KL}}(P \| Q).$$

Let  $S(x) = \Sigma^{1/2}x + \mu$ . For any coupling  $\pi$  of  $P'$  and  $\gamma$ ,  $(S \times S)_{\#}\pi$  is a coupling of  $P$  and  $Q$ , and

$$\begin{aligned} \int \|x - y\|^2 d((S \times S)_{\#}\pi) &= \int \|\Sigma^{1/2}(u - v)\|^2 d\pi(u, v) \\ &\leq \|\Sigma^{1/2}\|_{\text{op}}^2 \int \|u - v\|^2 d\pi(u, v), \end{aligned}$$

where  $\|\cdot\|_{\text{op}}$  some operator norm. Taking the infimum over couplings yields

$$W_2(P, Q) \leq \|\Sigma^{1/2}\|_{\text{op}} W_2(P', \gamma).$$

Hence,

$$W_2^2(P, Q) \leq \lambda_{\max}(\Sigma) W_2^2(P', \gamma).$$

Then, the Talagrand inequality [36, 45] implies

$$W_2^2(P', \gamma) \leq 2\text{D}_{\text{KL}}(P' \| \gamma).$$

Therefore, we obtain

$$\begin{aligned} W_2^2(P, Q) &\leq \lambda_{\max}(\Sigma) W_2^2(P', \gamma) \\ &\leq 2\lambda_{\max}(\Sigma) \text{D}_{\text{KL}}(P' \| \gamma) \\ &= 2\lambda_{\max}(\Sigma) \text{D}_{\text{KL}}(P \| Q). \end{aligned}$$

□

□

## I Proof of Corollary 2

By Theorem 1 of [46], a mechanism  $\mathcal{M}$  satisfies  $(\delta, \rho, \mathcal{D})$ -PAC privacy where

$$\text{D}_{\text{KL}}(\mathbf{1}_{\delta} \| \mathbf{1}_{\delta\rho}) \leq \text{MI}(X; \mathcal{M}(X)).$$

Thus,

$$\begin{aligned} R_{\text{KL}}^{\delta} &\geq \text{IntP}_{\text{KL}}(\mathcal{D}) - \inf_{P_W} \text{D}_{\text{KL}}(P_{X, \mathcal{M}(X)} \| P_X \otimes P_W) \\ &\geq \text{IntP}_{\text{KL}}(\mathcal{D}) - \text{MI}(X; \mathcal{M}(X)), \end{aligned}$$

where  $\text{IntP}_{\text{KL}}(\mathcal{D}) = -\text{D}_{\text{KL}}(\mathcal{D} \| \mathcal{U}) = \mathcal{H}(X) - \mathbf{V}$ , where  $\mathbf{V} = \log(|X|)$  if  $\mathcal{H}$  is Shannon entropy, and  $\mathbf{V} = \log(\int_X dx)$  if  $\mathcal{H}$  is differential entropy. Thus, we get  $R_{\mathcal{F}}^{\delta} \geq \mathcal{H}(X | \mathcal{M}(X)) - \mathbf{V}$ . □

## J Proof of Proposition 1

Recall that  $Z = \mathcal{M}(X) + B$  with  $B \sim \mathcal{N}(0, \Sigma_B)$  independent of  $X$ , where  $\mathcal{M}$  is a deterministic mechanism. Then, we have

$$\text{MI}(X; Z) = \mathcal{H}(Z) - \mathcal{H}(Z | X) = \mathcal{H}(Z) - \mathcal{H}(B).$$

Now consider the Gaussian surrogate distribution  $\tilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$ , where  $\mu_Z = \mu_{\mathcal{M}(X)}$  and  $\Sigma_Z = \Sigma_{\mathcal{M}(X)} + \Sigma_B$ . Its entropy is given by

$$\mathcal{H}(\tilde{Z}) = \frac{1}{2} \log \left[ (2\pi e)^d \det(\Sigma_Z) \right],$$

with  $\tilde{Z} \sim \tilde{Q}_{\mathcal{M}}$ , and similarly,  $\mathcal{H}(B) = \frac{1}{2} \log \left[ (2\pi e)^d \det(\Sigma_B) \right]$ . Hence,

$$\begin{aligned} \frac{1}{2} \log \det \left( I_d + \Sigma_{\mathcal{M}(X)} \Sigma_B^{-1} \right) &= \frac{1}{2} \log \left( \frac{\det(\Sigma_Z)}{\det(\Sigma_B)} \right) \\ &= \mathcal{H}(\tilde{Z}) - \mathcal{H}(B). \end{aligned}$$

So, we obtain

$$\begin{aligned} \text{Gap}_d &= \left[ \mathcal{H}(\tilde{Q}_{\mathcal{M}}) - \mathcal{H}(B) \right] - \left[ \mathcal{H}(Z) - \mathcal{H}(B) \right] \\ &= \mathcal{H}(\tilde{Z}) - \mathcal{H}(Z). \end{aligned}$$



Let  $\tilde{q}$  be the density function of  $\tilde{Q}_{\mathcal{M}}$ , and let  $p$  be the density function of  $P_{\mathcal{M},B}$ . Since  $\tilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$  is Gaussian, we have

$$\begin{aligned}\mathcal{H}(\tilde{Z}) &= -\log \tilde{q}(z) \\ &= \frac{d}{2} \log(2\pi) + \frac{1}{2} \log \det \Sigma_Z \frac{1}{2} (z - \mu_Z)^\top \Sigma_Z^{-1} (z - \mu_Z).\end{aligned}$$

Taking expectation under  $p$  yields

$$\begin{aligned}\mathcal{H}(Z, \tilde{Z}) &= \frac{d}{2} \log(2\pi) \\ &\quad + \frac{1}{2} \log \det \Sigma_Z + \frac{1}{2} \mathbb{E}_q \left[ (Z - \mu_Z)^\top \Sigma_Z^{-1} (Z - \mu_Z) \right],\end{aligned}$$

since  $\tilde{Z} \sim \tilde{Q}_{\mathcal{M}}$  matches  $Z \sim P_Z$  in mean and covariance, we have

$$\mathbb{E}_q \left[ (Z - \mu_Z)^\top \Sigma_Z^{-1} (Z - \mu_Z) \right] = \text{tr}(\Sigma_Z^{-1} \Sigma_Z) = \text{tr}(I_d) = d.$$

Thus,

$$\begin{aligned}\mathcal{H}(Z, \tilde{Z}) &= \frac{d}{2} \log(2\pi) + \frac{1}{2} \log \det \Sigma_Z + \frac{d}{2} \\ &= \mathcal{H}(\tilde{Z}).\end{aligned}$$

Therefore,

$$\begin{aligned}\text{D}_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}}) &= \mathcal{H}(Z, \tilde{Z}) - \mathcal{H}(Z) \\ &= \mathcal{H}(\tilde{Z}) - \mathcal{H}(Z) \\ &= \text{Gap}_d.\end{aligned}$$

Therefore,  $\text{Gap}_d = \text{D}_{\text{KL}}(P_{\mathcal{M},B} \| \tilde{Q}_{\mathcal{M}}) \geq 0$ , with equality if and only if  $P_{\mathcal{M},B} = \tilde{Q}_{\mathcal{M}}$ , i.e.,  $Z$  is exactly Gaussian with distribution  $\mathcal{N}(\mu_Z, \Sigma_Z)$ .  $\square$

## K Proof of Proposition 2

Since  $B \sim \mathcal{N}(0, \Sigma_B)$ , we have  $\mathbb{E}[\|B\|_2^2] = \text{tr}(\mathbb{E}[BB^\top]) = \text{tr}(\Sigma_B)$ . Hence, minimizing  $\mathbb{E}[\|B\|_2^2]$  over zero-mean Gaussian is equivalent to minimizing the trace  $\text{tr}(\Sigma_B)$  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  $\tilde{Q}_{\mathcal{M}} = \mathcal{N}(\mu_Z, \Sigma_Z)$  is the Gaussian distribution with the same first and second moments as  $Z$ . Then, by standard Gaussian-entropy formulas, we have

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

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

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

$$\Sigma_{\mathcal{M}(X)} = U \text{diag}(r_1, \dots, r_d) U^\top, \quad r_i > 0,$$

where  $U$  is the orthogonal eigenvector matrix from the eigen-decomposition of  $\Sigma_{\mathcal{M}(X)}$ . Writing  $\Sigma_B = \hat{U} \text{diag}(\ell_1, \dots, \ell_d) \hat{U}^\top$  with  $\ell_i > 0$ , the problem

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

becomes

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

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

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

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

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 have the following:

- As  $\lambda \rightarrow 0^+$ , each  $\ell_i(\lambda) \rightarrow 0^+$ , leading to  $F(\lambda) \rightarrow +\infty$ .
- As  $\lambda \rightarrow +\infty$ , each  $\ell_i(\lambda) \rightarrow +\infty$ , leading to  $F(\lambda) \rightarrow +\infty$ .

In addition,  $\frac{dF(\lambda)}{d\lambda} < 0$  throughout. Thus,  $F$  is strictly decreasing from  $+\infty$  down to 0. Therefore, there is a unique  $\lambda^* > 0$  such that  $F(\lambda^*) = \beta$ . At this  $\lambda^*$ , each  $\ell_i^* = \ell_i^*(\lambda^*)$  is unique. Thus,  $\Sigma_B^* = \hat{U} \text{diag}(\ell_1^*, \dots, \ell_d^*) \hat{U}^\top$  is unique minimizer of  $\text{tr}(\Sigma_B)$ . By construction,

$$\frac{1}{2} \log \det(I + \Sigma_{\mathcal{M}(X)} (\Sigma_B^*)^{-1}) = \beta.$$

Therefore, it is also the unique minimizer of (5).  $\square$

## L Proof of Proposition 3

For additive Gaussian noise with covariance  $\Sigma_B \succ 0$ ,

$$\text{MI}(X; \mathcal{M}(X) + B) \leq \frac{1}{2} \log \det(I + \Sigma_{\mathcal{M}(X)} \Sigma_B^{-1}).$$

The trace-optimal (eigen-aligned) choice that enforces the log-det constraint at level  $\hat{\beta}$  has eigenvalues  $e_i^* = \alpha\sqrt{\lambda_i}$  in the  $U$ -basis, with  $\alpha = \frac{S}{2\hat{\beta}}$ . Hence

$$\mathbb{E}\|B\|_2^2 = \text{tr}(\Sigma_B^*) = \sum_i e_i^* = \frac{S^2}{2\hat{\beta}}. \quad (1)$$

Algorithm 1 constructs a diagonal precision  $\Lambda_B$  in the empirical eigenbasis and, when the "gap" test passes, let

$$\lambda_{B,i} = \frac{2\nu}{\sqrt{\hat{\lambda}_i + \delta} \sum_k \sqrt{\hat{\lambda}_k + \delta}} \text{ and } \delta = \frac{10c\nu}{\beta}.$$

Plugging the population spectrum ( $\hat{\lambda} = \lambda$ , same  $U$ ) yields noise eigenvalues

$$e_i^\gamma = \frac{\sqrt{\lambda_i + \delta} \sum_k \sqrt{\lambda_k + \delta}}{2\nu},$$

and

$$\text{tr}(\Sigma_B^\gamma) = \frac{\left(\sum_k \sqrt{\lambda_k + \delta}\right)^2}{2\nu}.$$

Since  $\delta \geq 0$  and  $\nu \leq \hat{\beta}$ , we have

$$\text{tr}(\Sigma_B^\gamma) \geq \frac{S^2}{2\nu} \geq \frac{S^2}{2\hat{\beta}} = \text{tr}(\Sigma_B^*),$$

which proves (i) in this branch by (1)–(1). Moreover,

$$\frac{e_i^\gamma}{e_i^*} = \frac{\sum_k \sqrt{\lambda_k + \delta} \hat{\beta}}{\sum_k \sqrt{\lambda_k}} \frac{1}{\nu} \sqrt{1 + \frac{\delta}{\lambda_i}} \geq 1.$$

Thus,  $\Sigma_B^\gamma \succeq \Sigma_B^*$ . For additive Gaussian noise perturbation, increasing  $\Sigma_B$  (in positive semidefinite order) implies the log-det bound decreases, hence the mutual information decreases:

$$\begin{aligned} \text{MI}(X; \mathcal{M}(X) + B_\gamma) &\leq \frac{1}{2} \log \det (I + \Sigma_M(\Sigma_B^\gamma)^{-1}) \\ &\leq \frac{1}{2} \log \det (I + \Sigma_M(\Sigma_B^*)^{-1}). \end{aligned}$$

Since the Auto-PAC design saturates the bound at  $\hat{\beta}$  (and the true MI is bounded by it), we obtain

$$\text{MI}(X; \mathcal{M}(X) + B_\gamma) \leq \text{MI}(X; \mathcal{M}(X) + B),$$

establishing (ii) in this branch.

When the "gap" test fails, Algorithm 1 uses  $\Sigma_B^{\text{iso}} = \alpha I$  with  $\alpha = (\sum_i \hat{\lambda}_i + dc)/(2\nu)$ , hence

$$\text{tr}(\Sigma_B^{\text{iso}}) = \frac{d}{2\nu} \left( \sum_i \lambda_i + dc \right) \geq \frac{1}{2\nu} S^2 \geq \frac{S^2}{2\hat{\beta}} = \text{tr}(\Sigma_B^*),$$

where we used Cauchy–Schwarz  $S^2 = (\sum_i \sqrt{\lambda_i})^2 \leq d \sum_i \lambda_i$  and  $\nu \leq \hat{\beta}$ . Thus (i) also holds in this branch.

For (ii), both designs enforce the same budget  $\hat{\beta}$ :

$$\text{MI}(X; \mathcal{M}(X) + B_\gamma) \leq \hat{\beta}, \quad \text{MI}(X; \mathcal{M}(X) + B) \leq \hat{\beta}.$$

In the distinct-eigenvalues branch we proved the stronger order  $\leq$  between the two MI's. In the isotropic fallback, the same order holds whenever  $\alpha I \succeq \Sigma_B^*$  (e.g., for nearly isotropic spectra); otherwise we keep the common upper bound  $\hat{\beta}$ . Either way, the stated inequality (ii) is satisfied in the branch where Algorithm 1's eigenbasis matches  $U$ , and the confidence guarantee always preserves the budget.  $\square$

## M Proof of Proposition 4

Since  $\beta_1 < \beta_2$ , any distribution  $Q$  satisfying  $I_{\text{true}}(Q) \leq \beta_1$  necessarily satisfies  $I_{\text{true}}(Q) \leq \beta_2$ . Consequently, we have the inclusion  $\mathcal{F}(\beta_1) \subseteq \mathcal{F}(\beta_2)$ . Let  $A$  and  $\hat{A}$  be arbitrary sets with  $A \subseteq \hat{A}$ , 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  $f(Q) = \mathbb{E}_Q[\|B\|_2^2]$  yields

$$\inf_{Q \in \mathcal{F}(\beta_1)} \mathbb{E}_Q[\|B\|_2^2] \geq \inf_{Q \in \mathcal{F}(\beta_2)} \mathbb{E}_Q[\|B\|_2^2]$$

By definition,  $Q^*(\beta_i)$  achieves the infimum of  $\mathbb{E}_Q[\|B\|_2^2]$  over  $\mathcal{F}(\beta_i)$  for  $i = 1, 2$ . Therefore,

$$\begin{aligned} \mathbb{E}_{Q^*(\beta_1)}[\|B\|_2^2] &= \inf_{Q \in \mathcal{F}(\beta_1)} \mathbb{E}_Q[\|B\|_2^2] \geq \inf_{Q \in \mathcal{F}(\beta_2)} \mathbb{E}_Q[\|B\|_2^2] \\ &= \mathbb{E}_{Q^*(\beta_2)}[\|B\|_2^2]. \end{aligned}$$

$\square$

## N Proof of Theorem 2

By Lemma 3 (which is shown and proved later), the function  $g(\beta) = \text{Gap}_d(Q^*(\beta))$  is nondecreasing in  $\beta$ . Thus, for any  $0 < \beta_1 < \beta_2$ , we have  $\text{Gap}_d(Q^*(\beta_2)) \geq \text{Gap}_d(Q^*(\beta_1))$ , which yields  $G(\beta_2, \beta_1) = \text{Gap}_d(Q^*(\beta_2)) - \text{Gap}_d(Q^*(\beta_1)) \geq 0$ . Recall the relationship between true mutual information and the bound  $\text{LogDet}(\mathcal{M}(X), B) = \beta$ :

$$I_{\text{true}}(Q^*(\beta)) = \beta - \text{Gap}_d(Q^*(\beta)).$$

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

$$\begin{aligned} I_{\text{true}}(Q^*(\beta_2)) - I_{\text{true}}(Q^*(\beta_1)) &= [\beta_2 - \text{Gap}_d(Q^*(\beta_2))] - [\beta_1 - \text{Gap}_d(Q^*(\beta_1))] \\ &= (\beta_2 - \beta_1) - [\text{Gap}_d(Q^*(\beta_2)) - \text{Gap}_d(Q^*(\beta_1))] \\ &= (\beta_2 - \beta_1) - G(\beta_2, \beta_1). \end{aligned}$$

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

$$I_{\text{true}}(Q^*(\beta_2)) - I_{\text{true}}(Q^*(\beta_1)) = (\beta_2 - \beta_1) - G(\beta_2, \beta_1) \geq 0,$$

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

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

i.e.  $\mathbb{I}_{\text{true}}(Q^*(\beta_1)) > \mathbb{I}_{\text{true}}(Q^*(\beta_2))$ .  $\square$

**Lemma 3.** Fix a mechanism  $\mathcal{M}$  and a data distribution  $\mathcal{D}$ . Let  $Q^*(\beta)$  be the solution of (5). Then,  $\text{Gap}_d(Q^*(\beta))$  is a nondecreasing function of  $\beta$ .

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

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

Differentiate with respect to  $\beta$  via the chain rule:

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

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 [44],

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

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,

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

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

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

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

## O Proof of Proposition 5

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

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

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

$$\begin{aligned} W(Q, \pi) - W(Q, \eta_Q) &= \int_{\mathcal{X}, \mathcal{Y}, \mathbb{R}^d} P_X(x) G_{\mathcal{M}, Q}(y|x, b) \log \eta_Q(x|y+b) dx dy db \\ &\quad - \int_{\mathcal{X}, \mathcal{Y}, \mathbb{R}^d} P_X(x) G_{\mathcal{M}, Q}(y|x, b) \log \pi(x|y+b) dx dy db \\ &= \int_{\mathcal{X}, \mathcal{Y}, \mathbb{R}^d} P_X(x) G_{\mathcal{M}, Q}(y|x, b) \log \frac{\eta_Q(x|y+b)}{\pi(x|y+b)} dx dy db. \end{aligned}$$

Let

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

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  $Q \in \Gamma$ ,

$$W(Q, \pi) - W(Q, \eta_Q) = D_{\text{KL}}(\eta_Q \| \pi) \geq 0.$$

Then,  $W(Q, \pi) \geq W(Q, \eta_Q)$ , 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.  $\square$

## P Proof of Proposition 6

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

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

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

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

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

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

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

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

## Q Proof of Proposition 7

### Q.1 Part (i):

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

$$\mathcal{H}(Z + B_{\text{pac}}) < \mathcal{H}(Z_G + B_{\text{pac}}),$$

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

### Q.2 Part (ii):

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

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

For the stationarity condition w.r.t. each  $\sigma_w^2$  we need the gradient of mutual information. By [37], we have

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

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

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

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

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

these equalities can hold only for as long as the required mutual information reduction does not exceed  $\beta_{1\text{ab}}$ . Therefore,  $\sigma_v^2 = 0$  for every  $v \in S_{1\text{ab}}$ . With those label-directions undisturbed, each class margin  $e_\ell - e_j$  retains its sign, whence  $\arg \max_i (Z_i + B_i^*) = \hat{y}$ .  $\square$

## R Proof of Theorem 5

### R.1 Part (i)

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

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

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

where  $\text{Gap}_d(\beta) = D_{\text{KL}}(P_{\mathcal{M}, B_{\text{PAC}}} \| \tilde{Q}_{\mathcal{M}}) \geq 0$ . When  $\mathcal{M}(X)$  is non-Gaussian,  $\text{Gap}_d(\beta) > 0$  for all  $\beta > 0$ . By de Bruijn's identity (e.g., [38]),

$$\frac{d}{d\beta} \text{Gap}_d(\beta) = \frac{1}{2} \mathcal{J}(P_{\mathcal{M} + B_{\text{PAC}}}(\beta) \| \tilde{Q}_{\mathcal{M}}) > 0,$$

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

### R.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  $Q \mapsto \text{MI}(Q) \equiv \text{MI}(X; \mathcal{M}(X) + B)$  is convex. The objective  $\mathcal{K}(Q) = \mathbb{E}_Q[\|B\|_2^2]$  is linear (hence convex) in  $Q$ . In addition, the constraint set  $\{Q : \text{MI}(Q) \leq \beta\}$  is convex. Then, Slater's condition holds because:

(i) when  $\Sigma_B \rightarrow \infty$ ,  $\text{MI}(Q) \rightarrow 0 < \beta$ ;

(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

$$\hat{V}(\beta) = \max_Q \mathcal{K}(Q) + \lambda(\text{MI}(Q) - \beta).$$

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

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

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

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

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



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

By Lemma 4 (stated and proved below) to  $g(\beta) = V_{\text{PAC}}(\beta)$  and  $f(\beta) = V(\beta)$ , we have  $g'(\beta) > f'(\beta)$  for all  $\beta > 0$ . That is,  $V'_{\text{PAC}}(\beta) > V'_{\text{SR}}(\beta)$ . Thus,  $\text{Util}_{\beta}^{\text{SR}} \geq \text{Util}_{\beta}^{\text{PAC}}$ , with equality only for Gaussian  $\mathcal{M}(X)$ .  $\square$

**Lemma 4** (Height gap  $\Rightarrow$  slope gap). *Let  $g, f : (0, \infty) \rightarrow \mathbb{R}$  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  $g'(\beta) > f'(\beta)$ .  $\square$

## S More on Experiments

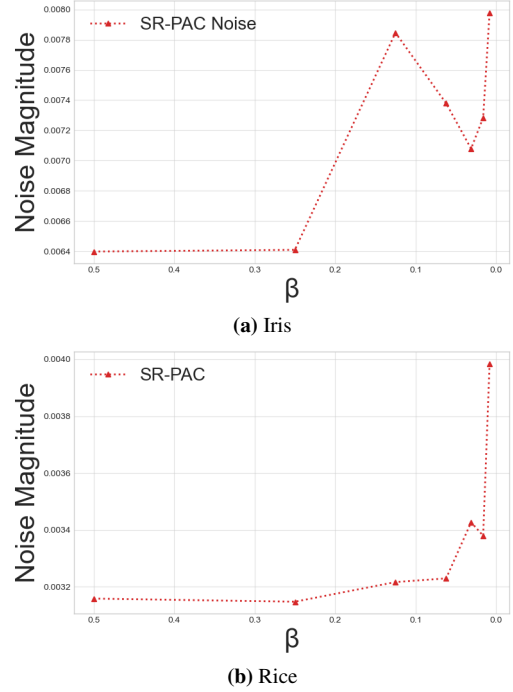
### S.1 More On The results of Fig. 2

Fig. 3 zooms in on the SR-PAC curves from Fig. 2 and shows a clear monotone increase in expected noise power  $\mathbb{E}\|B\|^2 = \text{tr}(\Sigma)$  as  $\beta$  decreases. This is consistent with the tighter privacy requirement  $H_c \geq H_M - \beta$  pushing the mechanism to add more noise in the high-privacy (small- $\beta$ ) regime.

Fig. 4 further shows that SR-PAC implements the privacy constraint *conservatively*: the achieved conditional entropy  $H_c$  typically lies at or slightly above the target line  $\mathcal{H}(X) - \beta$ . Equivalently, the effective mutual information  $\text{MI}(M; Y) = \mathcal{H}(X) - H_c$  is at or below the nominal budget  $\beta$ ; i.e., the realized mechanism is (slightly) *more private than necessary*. This benign overshoot is expected from finite-sample estimation and our fixed-CRN calibration with a positive tolerance, which is designed to avoid budget violations. If desired, the conservatism can be reduced by tightening the calibration tolerance, enlarging the CRN bank, or applying a final back-off on the noise scale until  $H_c$  falls within a small band above the target. Importantly, even with this conservative bias, SR-PAC attains lower noise power and higher accuracy than Auto-PAC and Efficient-PAC at the same nominal  $\beta$ .

### S.2 Empirical Membership Inference Attack

We use the Likelihood-Ratio Attack (LIRA) described in [6] to perform the empirical membership inference attacks (MIAs) on the mechanisms privatized by SR-PAC, Auto-PAC, Efficient-PAC, and DP in Section 6.2, using Iris and Rice datasets. The empirical posterior success rate (PSR) is measured as the average accuracy of the MIA.

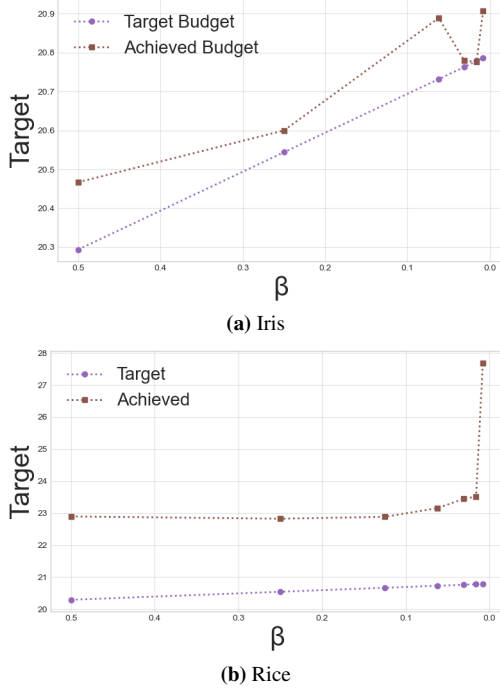


**Figure 3:** Noise magnitudes of SR-PAC of Fig. 2 a and b. All the numerical values are shown in Tables 5 and 6.

A theoretical ordering of privacy budgets in mutual information or conditional entropy does not in general imply an ordering of a membership-inference attack’s PSR. Finite-sample effects, non-optimal attacks, calibration error, and run-to-run variance can all break that implication. We therefore read the PSR curves in Fig. 5 empirically, as a *diagnostic* rather than a ground-truth ranking of privacy strength.

As  $\beta$  decreases (higher privacy), all mechanisms tend to push PSR toward 0.5 (chance), but the trends are not strictly monotone and the rankings cross. This is expected. In particular, DP is generally conservative and can display lower PSR at very small  $\beta$ , while PAC-based methods may sit closer to chance but with visible fluctuations. Auto-PAC and Efficient-PAC allocate anisotropic noise from second-order structure (covariance scaling or eigen-allocation); in small-sample regimes (e.g. Iris and Rice), those moment estimates are noisy or ill-conditioned, so the resulting implementations are unstable and can become conservative (over-noisy), which may depress PSR at a fixed  $\beta$ .

Our goal with SR-PAC is privacy budget fidelity (i.e., to address the conservativeness of Auto-PAC and Efficient-PAC privatization), not to minimize PSR per se. SR-PAC enforces the conditional-entropy target directly and typically attains the desired leakage with less noise than Auto-PAC and Efficient-PAC. Hence it is plausible—and observed in Fig. 5—that SR-PAC’s PSR can be comparable to, or occasionally above, over-noised baselines at the same  $\beta$ . The take-away is that PSR complements our main metric: SR-PAC achieves tighter



**Figure 4:** SR-PAC’s performance of implementing the target privacy budget of Fig. 2 a and b. All the numerical values are shown in Tables 5 and 6.

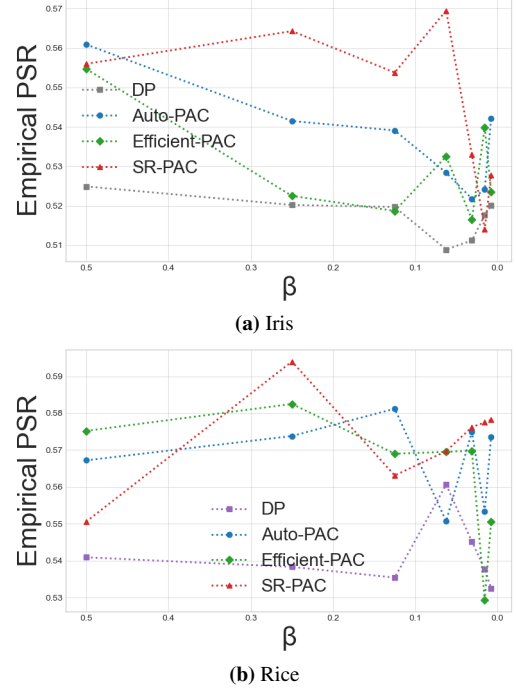
budget implementation with lower noise power, while differences in PSR reflect both budget alignment and attack/model mismatch rather than a strict ordering induced by  $\beta$ .

### S.3 Discussion On The results of Fig. 1

Fig. 1 empirically demonstrates that SR-PAC enforces the mutual-information budget more tightly than Auto-PAC and Efficient-PAC, yielding higher accuracy with lower noise. The performance gaps (both accuracy and noise magnitude) increase as  $\beta$  decreases (i.e., in higher-privacy regimes). This behavior can be understood as follows.

*Budget alignment vs. Gaussian surrogate.* SR-PAC enforces the privacy constraint directly in terms of conditional entropy, aligning the mutual information budget with the actual leakage bound. In contrast, Auto-PAC and Efficient-PAC rely on Gaussian surrogates that ignore the higher-order, non-Gaussian structure of the outputs, leading to conservative privacy budget implementation. This conservativeness can lead them to add more noise than necessary to meet a given  $\beta$ , and the inefficiency becomes a larger fraction of the total budget when  $\beta$  is small, amplifying their disadvantage in high-privacy regimes.

*Directional selectivity under tight budgets.* SR-PAC learns an anisotropic, task-directed noise shape that concentrates perturbations away from task-critical directions (Theorem 7), thereby preserving classification margins while still raising



**Figure 5:** The performance of empirical membership inference attack using empirical LIRA, measured by the empirical posterior success rate (PSR). All the numerical values are shown in Tables 5 and 6.

conditional entropy to the desired level. By contrast, Auto-PAC and Efficient-PAC are also anisotropic but task-agnostic: Auto-PAC scales the raw logit covariance ( $\Sigma_B \propto \Sigma_{\text{raw}}$ ), and Efficient-PAC allocates along the eigenbasis with  $e_i \propto \sqrt{\lambda_i}$ . Both rely only on second-order statistics and ignore label-conditioned and higher-order structure, so they can spend budget along decision-sensitive axes whenever those coincide with high-variance directions. This mismatch is particularly costly under tight privacy budgets (small  $\beta$ ), where misallocated power yields larger accuracy loss for the same leakage target.

*Non-Gaussian exploitation.* As  $\beta$  decreases, the non-Gaussian structure of the outputs matters more. SR-PAC can use less total noise by optimally exploiting the geometries of the outputs (e.g., via leveraging flexible posteriors and calibration with fixed common random numbers). Gaussian surrogates cannot capture this effect, so their "privacy per unit noise" degrades as  $\beta$  shrinks.

*Utility sensitivity.* Viewing the required noise power as a utility curve  $\text{Util}_\beta^{\text{SR}}$  (Theorem 5), SR-PAC exhibits better sensitivity (i.e., larger accuracy retention per unit budget). As  $\beta$  decreases, the noise power of Auto-PAC and Efficient-PAC baselines grows faster than that of SR-PAC, widening the gap in both accuracy and magnitude. When  $\beta$  is large (loose privacy), all methods add little noise and their performance converges. As  $\beta$  decreases (stricter privacy), SR-PAC

optimally places noise where it is least harmful, yielding progressively larger gains over Auto-PAC and Efficient-PAC in both accuracy and noise efficiency.

**Table 1:** CIFAR-10 dataset results. Accuracy: higher is better. Noise Magnitude  $\mathbb{E}[\|B\|_2^2]$ : lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Accuracy (Fig. 1a)				(b) Noise Magnitude (Fig. 1e)				(c) Target Match (Fig. 1i)				
$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Target	Achieved	$\Delta$	Rel.%
10.00	64.3 %	62.8 %	<b>67.7%</b>	10.00	203.6	194.1	<b>86.717</b>	10.00	3.0	<b>2.9762</b>	0.0238	0.8
9.50	63.5 %	61.8 %	<b>67.2%</b>	9.50	228.8	218.3	<b>101.123</b>	9.50	3.5	<b>3.4901</b>	0.0099	0.3
9.00	62.8 %	60.8 %	<b>66.7%</b>	9.00	257.6	245.9	<b>113.220</b>	9.00	4.0	<b>3.9872</b>	0.0128	0.3
8.50	61.8 %	59.9 %	<b>66.0%</b>	8.50	290.8	277.8	<b>127.892</b>	8.50	4.5	<b>4.4889</b>	0.0111	0.3
8.00	61.0 %	58.8 %	<b>65.2%</b>	8.00	329.1	314.6	<b>144.288</b>	8.00	5.0	<b>4.9873</b>	0.0127	0.3
7.50	59.7 %	57.8 %	<b>64.3%</b>	7.50	373.6	357.5	<b>162.920</b>	7.50	5.5	<b>5.4868</b>	0.0132	0.2
7.00	58.5 %	56.7 %	<b>63.5%</b>	7.00	425.8	407.7	<b>184.447</b>	7.00	6.0	<b>5.9853</b>	0.0147	0.2
6.50	57.0 %	55.5 %	<b>62.6%</b>	6.50	487.4	467.1	<b>209.897</b>	6.50	6.5	<b>6.4895</b>	0.0105	0.2
6.00	55.5 %	54.1 %	<b>61.3%</b>	6.00	560.7	537.8	<b>238.969</b>	6.00	7.0	<b>6.9897</b>	0.0103	0.2
5.50	53.9 %	52.5 %	<b>60.0%</b>	5.50	649.1	623.2	<b>272.759</b>	5.50	7.5	<b>7.4881</b>	0.0119	0.2
5.00	52.1 %	50.9 %	<b>58.7%</b>	5.00	757.1	727.6	<b>311.272</b>	5.00	8.0	<b>7.9863</b>	0.0137	0.2
4.50	50.0 %	49.1 %	<b>57.0%</b>	4.50	891.3	857.5	<b>357.395</b>	4.50	8.5	<b>8.4852</b>	0.0148	0.2
4.00	48.0 %	46.8 %	<b>55.0%</b>	4.00	1061.5	1022.4	<b>409.579</b>	4.00	9.0	<b>8.9867</b>	0.0133	0.2
3.50	45.6 %	44.2 %	<b>53.1%</b>	3.50	1283.3	1237.3	<b>474.419</b>	3.50	9.5	<b>9.4868</b>	0.0132	0.1
3.00	42.6 %	41.8 %	<b>50.9%</b>	3.00	1582.4	1527.5	<b>553.758</b>	3.00	10.0	<b>9.9873</b>	0.0127	0.1
2.50	39.3 %	38.6 %	<b>48.2%</b>	2.50	2005.3	1938.1	<b>652.985</b>	2.50	10.5	<b>10.4908</b>	0.0092	0.1
2.00	35.8 %	35.2 %	<b>45.1%</b>	2.00	2645.0	2559.4	<b>781.609</b>	2.00	11.0	<b>10.9896</b>	0.0104	0.1
1.50	31.5 %	31.3 %	<b>41.4%</b>	1.50	3718.4	3602.3	<b>979.469</b>	1.50	11.5	<b>11.4915</b>	0.0085	0.1
1.00	26.8 %	26.8 %	<b>36.4%</b>	1.00	5875.7	5699.4	<b>1290.274</b>	1.00	12.0	<b>11.9920</b>	0.0080	0.1
0.50	20.8 %	20.9 %	<b>30.9%</b>	0.50	12369.4	12013.0	<b>1844.353</b>	0.50	12.5	<b>12.4939</b>	0.0061	0.1
0.25	17.0 %	17.1 %	<b>27.5%</b>	0.25	25373.0	24657.0	<b>2327.319</b>	0.25	12.8	<b>12.7420</b>	0.0080	0.1

**Table 2:** CIFAR-100 dataset results. Accuracy: higher is better. Noise Magnitude  $\mathbb{E}[\|B\|_2^2]$ : lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Accuracy (Fig. 1b)				(b) Noise Magnitude (Fig. 1e)				(c) Target Match (Fig. 1i)				
$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Target	Achieved	$\Delta$	Rel.%
80.00	55.7%	56.4%	<b>59.1%</b>	80.00	295.7	156.1	<b>0.015</b>	80.00	30.0	<b>32.9105</b>	2.9105	9.7
75.00	55.2%	56.2%	<b>59.3%</b>	75.00	335.7	166.5	<b>9.210</b>	75.00	35.0	<b>35.0027</b>	0.0027	0.0
70.00	54.5%	56.0%	<b>58.4%</b>	70.00	382.6	178.4	<b>73.285</b>	70.00	40.0	<b>40.0269</b>	0.0269	0.1
65.00	53.8%	55.7%	<b>57.8%</b>	65.00	437.9	192.1	<b>121.950</b>	65.00	45.0	<b>45.0010</b>	0.0010	0.0
60.00	52.8%	55.4%	<b>57.2%</b>	60.00	503.8	208.1	<b>173.816</b>	60.00	50.0	<b>50.0244</b>	0.0244	0.1
55.00	51.8%	55.1%	<b>56.2%</b>	55.00	583.3	227.0	<b>217.047</b>	55.00	55.0	<b>55.0116</b>	0.0116	0.0
50.00	50.4%	54.6%	<b>55.7%</b>	50.00	680.3	249.7	<b>256.837</b>	50.00	60.0	<b>60.0306</b>	0.0306	0.1
45.00	48.9%	54.1%	<b>55.8%</b>	45.00	800.9	277.5	<b>267.036</b>	45.00	65.0	<b>65.0453</b>	0.0453	0.1
40.00	46.9%	53.7%	<b>54.8%</b>	40.00	953.8	312.2	<b>318.676</b>	40.00	70.0	<b>70.0253</b>	0.0253	0.0
35.00	44.4%	52.9%	<b>53.4%</b>	35.00	1153.1	356.8	<b>409.873</b>	35.00	75.0	<b>75.0114</b>	0.0114	0.0
30.00	41.3%	51.6%	<b>52.1%</b>	30.00	1421.9	416.2	<b>480.533</b>	30.00	80.0	<b>80.0208</b>	0.0208	0.0
25.00	37.6%	50.1%	<b>50.3%</b>	25.00	1801.9	499.5	<b>659.901</b>	25.00	85.0	<b>85.0612</b>	0.0612	0.1
20.00	33.3%	48.3%	<b>49.2%</b>	20.00	2376.8	624.3	<b>709.042</b>	20.00	90.0	<b>90.0738</b>	0.0738	0.1
15.00	27.5%	45.4%	<b>48.0%</b>	15.00	3341.2	832.4	<b>799.096</b>	15.00	95.0	<b>95.0488</b>	0.0488	0.1
10.00	20.4%	39.8%	<b>46.6%</b>	10.00	5279.8	1248.6	<b>927.919</b>	10.00	100.0	<b>100.0729</b>	0.0729	0.1
7.00	15.2%	34.2%	<b>46.3%</b>	7.00	7778.9	1783.8	<b>955.096</b>	7.00	103.0	<b>103.0800</b>	0.0800	0.1
5.00	11.7%	28.7%	<b>45.5%</b>	5.00	11114.8	2497.3	<b>1060.981</b>	5.00	105.0	<b>105.0304</b>	0.0304	0.0
4.00	9.6%	24.7%	<b>45.0%</b>	4.00	14035.3	3121.6	<b>1027.777</b>	4.00	106.0	<b>106.1096</b>	0.1096	0.1
3.00	7.5%	20.1%	<b>44.1%</b>	3.00	18904.0	4162.1	<b>1057.745</b>	3.00	107.0	<b>107.0857</b>	0.0857	0.1
2.00	5.5%	15.0%	<b>44.8%</b>	2.00	28643.4	6243.2	<b>1114.597</b>	2.00	108.0	<b>108.0618</b>	0.0618	0.1
1.00	3.3%	8.7%	<b>45.0%</b>	1.00	57865.4	12486.4	<b>1059.941</b>	1.00	109.0	<b>109.0935</b>	0.0935	0.1
0.50	2.3%	5.1%	<b>45.1%</b>	0.50	116312.4	24972.7	<b>1062.913</b>	0.50	109.5	<b>109.5334</b>	0.0334	0.0

**Table 3:** MNIST dataset. Accuracy: higher is better. Noise Magnitude  $\mathbb{E}[\|B\|_2^2]$ : lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Accuracy (Fig. 1c)				(b) Noise Magnitude (Fig. 1g)				(c) Target Match (Fig. 1k)				
$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Target	Achieved	$\Delta$	Rel.%
7.00	92.3%	85.6%	<b>98.4%</b>	7.00	86.7	133.5	<b>0.000</b>	7.00	4.7	<b>5.8283</b>	1.0805	22.8
6.50	91.3%	84.3%	<b>98.4%</b>	6.50	99.3	143.7	<b>0.000</b>	6.50	5.2	<b>5.8283</b>	0.5805	11.1
6.00	89.4%	82.9%	<b>98.4%</b>	6.00	114.2	155.7	<b>0.000</b>	6.00	5.7	<b>5.8283</b>	0.0805	1.4
5.50	87.1%	81.0%	<b>97.0%</b>	5.50	132.2	169.9	<b>28.292</b>	5.50	6.2	<b>6.2481</b>	0.0003	0.0
5.00	84.4%	78.9%	<b>95.4%</b>	5.00	154.2	186.8	<b>55.036</b>	5.00	6.7	<b>6.7489</b>	0.0011	0.0
4.50	81.1%	76.5%	<b>93.4%</b>	4.50	181.6	207.6	<b>77.870</b>	4.50	7.2	<b>7.2529</b>	0.0051	0.1
4.00	77.5%	73.8%	<b>91.3%</b>	4.00	216.2	233.6	<b>97.935</b>	4.00	7.7	<b>7.7541</b>	0.0063	0.1
3.50	73.2%	70.4%	<b>89.1%</b>	3.50	261.4	266.9	<b>116.500</b>	3.50	8.2	<b>8.2578</b>	0.0100	0.1
3.00	68.0%	66.5%	<b>87.1%</b>	3.00	322.4	311.4	<b>133.862</b>	3.00	8.7	<b>8.7573</b>	0.0095	0.1
2.50	62.1%	62.0%	<b>85.2%</b>	2.50	408.5	373.7	<b>150.256</b>	2.50	9.2	<b>9.2510</b>	0.0032	0.0
2.00	55.3%	56.7%	<b>83.6%</b>	2.00	538.8	467.1	<b>164.749</b>	2.00	9.7	<b>9.7523</b>	0.0045	0.1
1.50	47.3%	50.2%	<b>82.0%</b>	1.50	757.5	622.8	<b>179.350</b>	1.50	10.2	<b>10.2496</b>	0.0018	0.0
1.00	38.5%	41.9%	<b>80.5%</b>	1.00	1197.0	934.2	<b>194.086</b>	1.00	10.7	<b>10.7519</b>	0.0041	0.0
0.50	28.1%	30.6%	<b>79.1%</b>	0.50	2519.9	1868.4	<b>208.289</b>	0.50	11.2	<b>11.2606</b>	0.0128	0.1
0.25	21.0%	23.5%	<b>78.3%</b>	0.25	5169.0	3736.8	<b>215.495</b>	0.25	11.5	<b>11.5088</b>	0.0110	0.1



**Table 4:** Ag-news dataset results. Accuracy: higher is better. Noise Magnitude  $\mathbb{E}[\|B\|_2^2]$ : lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Accuracy (Fig. 1d)				(b) Noise Magnitude (Fig. 1h)				(c) Target Match (Fig. 1l)				
$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Auto-PAC	Efficient-PAC	SR-PAC	$\beta$	Target	Achieved	$\Delta$	Rel.%
9.50	96.7%	89.6%	<b>96.6%</b>	9.50	1.2	20.7	<b>0.54</b>	9.50	0.2	<b>0.2112</b>	0.0003	0.1
9.00	96.6%	89.3%	<b>95.8%</b>	9.00	1.5	21.9	<b>1.79</b>	9.00	0.7	<b>0.7119</b>	0.0010	0.1
8.50	96.5%	88.9%	<b>95.1%</b>	8.50	2.0	23.2	<b>3.07</b>	8.50	1.2	<b>1.2166</b>	0.0057	0.5
8.00	96.3%	88.4%	<b>94.6%</b>	8.00	2.5	24.6	<b>3.99</b>	8.00	1.7	<b>1.7179</b>	0.0070	0.4
7.50	96.0%	87.9%	<b>93.5%</b>	7.50	3.3	26.2	<b>6.05</b>	7.50	2.2	<b>2.2207</b>	0.0098	0.4
7.00	95.7%	87.3%	<b>92.6%</b>	7.00	4.3	28.1	<b>7.70</b>	7.00	2.7	<b>2.7178</b>	0.0069	0.3
6.50	95.3%	86.7%	<b>92.3%</b>	6.50	5.5	30.3	<b>8.70</b>	6.50	3.2	<b>3.2152</b>	0.0043	0.1
6.00	94.8%	85.9%	<b>91.8%</b>	6.00	7.2	32.8	<b>9.90</b>	6.00	3.7	<b>3.7220</b>	0.0111	0.3
5.50	94.1%	85.1%	<b>91.6%</b>	5.50	9.3	35.8	<b>10.57</b>	5.50	4.2	<b>4.2152</b>	0.0043	0.1
5.00	93.1%	84.2%	<b>91.6%</b>	5.00	12.2	39.4	<b>10.24</b>	5.00	4.7	<b>4.7198</b>	0.0089	0.2
4.50	91.9%	83.0%	<b>89.3%</b>	4.50	16.1	43.7	<b>15.72</b>	4.50	5.2	<b>5.2149</b>	0.0040	0.1
4.00	90.2%	81.7%	<b>89.5%</b>	4.00	21.4	49.2	<b>15.88</b>	4.00	5.7	<b>5.7188</b>	0.0079	0.1
3.50	88.1%	80.1%	<b>89.5%</b>	3.50	28.7	56.2	<b>16.05</b>	3.50	6.2	<b>6.2142</b>	0.0033	0.1
3.00	85.2%	78.1%	<b>89.4%</b>	3.00	39.2	65.6	<b>16.49</b>	3.00	6.7	<b>6.7129</b>	0.0020	0.0
2.50	81.5%	75.6%	<b>87.1%</b>	2.50	54.8	78.7	<b>23.51</b>	2.50	7.2	<b>7.2273</b>	0.0164	0.2
2.00	76.6%	72.4%	<b>88.6%</b>	2.00	79.5	98.4	<b>19.39</b>	2.00	7.7	<b>7.7260</b>	0.0151	0.2
1.50	70.2%	68.0%	<b>86.2%</b>	1.50	122.3	131.2	<b>25.96</b>	1.50	8.2	<b>8.2306</b>	0.0197	0.2
1.00	61.8%	61.8%	<b>87.8%</b>	1.00	210.5	196.8	<b>21.14</b>	1.00	8.7	<b>8.7154</b>	0.0045	0.1
0.50	50.2%	52.0%	<b>87.0%</b>	0.50	480.8	393.6	<b>23.16</b>	0.50	9.2	<b>9.2115</b>	0.0006	0.0
0.25	42.2%	44.2%	<b>87.6%</b>	0.25	1025.7	787.3	<b>21.77</b>	0.25	9.5	<b>9.4754</b>	0.0145	0.2
0.08	35.4%	36.9%	<b>85.1%</b>	0.08	3346.4	2460.3	<b>26.33</b>	0.08	9.6	<b>9.6524</b>	0.0215	0.2
0.06	34.2%	35.6%	<b>86.0%</b>	0.06	4484.3	3280.3	<b>27.45</b>	0.06	9.7	<b>9.6534</b>	0.0025	0.0
0.02	33.6%	34.9%	<b>86.8%</b>	0.02	13588.6	9841.0	<b>29.47</b>	0.02	9.7	<b>9.6975</b>	0.0066	0.1

**Table 5:** Iris dataset results. Empirical Posterior Success Rate (PSR), Noise Magnitude, and Target Match. Empirical PSR: higher is better. Noise Magnitude: lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Empirical PSR					(b) Noise Magnitude					(c) Target Match (SR-PAC)				
$\beta$	SR-PAC	DP	Auto-PAC	Efficient-PAC	$\beta$	SR-PAC	DP	Auto-PAC	Efficient-PAC	$\beta$	Target $H_c$	Achieved $H_c$	$\Delta$	Rel.%
0.5000	<b>0.5560</b>	0.5249	0.5609	0.5547	0.5000	<b>0.006399</b>	0.0042	0.0865	0.0256	0.5000	20.3	<b>20.467</b>	0.1730	0.9
0.2500	<b>0.5643</b>	0.5202	0.5415	0.5225	0.2500	<b>0.006410</b>	0.0085	0.0909	0.0261	0.2500	20.5	<b>20.600</b>	0.0560	0.3
0.1250	<b>0.5538</b>	0.5197	0.5391	0.5187	0.1250	<b>0.007846</b>	0.0136	0.1073	0.0321	0.1250	20.7	<b>20.884</b>	0.2150	1.0
0.0625	<b>0.5695</b>	0.5089	0.5284	0.5325	0.0620	<b>0.007379</b>	0.0208	0.0991	0.0283	0.0620	20.7	<b>20.889</b>	0.1570	0.8
0.0312	<b>0.5330</b>	0.5112	0.5218	0.5165	0.0310	<b>0.007079</b>	0.0314	0.1087	0.0400	0.0310	20.8	<b>20.780</b>	0.0170	0.1
0.0156	<b>0.5141</b>	0.5176	0.5241	0.5399	0.0160	<b>0.007284</b>	0.0481	0.1147	0.0379	0.0160	20.8	<b>20.776</b>	-0.0030	0.0
0.0078	<b>0.5278</b>	0.5201	0.5422	0.5236	0.0080	<b>0.007979</b>	0.0753	0.1161	0.0381	0.0080	20.8	<b>20.908</b>	0.1210	0.6

**Table 6:** Rice dataset results. Empirical Posterior Success Rate (PSR), Noise Magnitude, and Target Match. Empirical PSR: higher is better. Noise Magnitude: lower is better. Target Match: smaller  $\Delta/\text{Rel.}\%$  is better. All results are averaged over 35 trials.

(a) Empirical PSR					(b) Noise Magnitude					(c) Target Match (SR-PAC)				
$\beta$	SR-PAC	DP	Auto-PAC	Efficient-PAC	$\beta$	SR-PAC	DP	Auto-PAC	Efficient-PAC	$\beta$	Target	Achieved	$\Delta$	Rel.%
0.5000	<b>0.5506</b>	0.5409	0.5672	0.5752	0.5000	<b>0.003159</b>	0.0020	0.1168	0.0112	0.5000	20.3	<b>22.896</b>	2.6020	12.8
0.2500	<b>0.5940</b>	0.5383	0.5738	0.5825	0.2500	<b>0.003148</b>	0.0030	0.1291	0.0117	0.2500	20.5	<b>22.825</b>	2.2810	11.1
0.1250	<b>0.5631</b>	0.5354	0.5812	0.5690	0.1250	<b>0.003217</b>	0.0038	0.1417	0.0165	0.1250	20.7	<b>22.884</b>	2.2150	10.7
0.0625	<b>0.5697</b>	0.5606	0.5507	0.5695	0.0620	<b>0.003230</b>	0.0047	0.1348	0.0115	0.0620	20.7	<b>23.156</b>	2.4240	11.7
0.0312	<b>0.5762</b>	0.5452	0.5749	0.5698	0.0310	<b>0.003427</b>	0.0056	0.1589	0.0225	0.0310	20.8	<b>23.444</b>	2.6810	12.9
0.0156	<b>0.5775</b>	0.5377	0.5534	0.5294	0.0160	<b>0.003380</b>	0.0068	0.2287	0.0475	0.0160	20.8	<b>23.516</b>	2.7370	13.2
0.0078	<b>0.5782</b>	0.5325	0.5736	0.5505	0.0080	<b>0.003985</b>	0.0082	0.2136	0.0907	0.0080	20.8	<b>27.687</b>	6.9000	33.2