JACOBIAN-GUIDED ANISOTROPIC NOISE RESHAPING FOR ENHANCING REPRESENTATION UTILITY UNDER LOCAL DIFFERENTIAL PRIVACY

Anonymous authors
Paper under double-blind review

ABSTRACT

While Local Differential Privacy (LDP) serves as a foundational primitive for distributed data collection, its stringent randomization requirements often lead to severe degradation in data representation utility. This degradation stems from the task-agnostic nature of conventional LDP mechanisms, which perturb all dimensions without accounting for their relative importance to the downstream objective. To address this issue, we propose a novel approach that mitigates noise in task-relevant subspaces of the data representation. Our method identifies task-critical subspaces via the Jacobian of a pre-trained downstream model, selectively attenuates noise along these directions, and reshapes the isotropic noise of standard LDP mechanisms into an anisotropic distribution. The resulting mechanism preserves the privacy guarantee of the underlying LDP randomizer while heterogeneously modulating the impact of noise across task directions, thereby substantially enhancing data utility. The approach is applicable to both linear and nonlinear models and can be seamlessly integrated with existing LDP mechanisms. Extensive experiments on CIFAR-10-C under brightness corruption at the highest severity level demonstrate that integrating our approach improves classification accuracy by approximately 8 % for Laplace and 20 % for PrivUnit variants at $\epsilon = 7.5$.

1 INTRODUCTION

Local Differential Privacy (LDP) (Kasiviswanathan et al., 2011) has become a foundational primitive for privacy-preserving distributed data collection. Unlike centralized approaches (Dwork et al., 2006), LDP eliminates the need for a trusted data curator by enabling each data owner to locally randomize their data prior to sharing. Owing to this decentralized trust model, it has seen widespread adoption in practice. Prominent examples include Google's RAPPOR for collecting statistics from Chrome clients (Erlingsson et al., 2014), Apple's deployment in iOS for collecting new word suggestions, emoji usage frequencies, and health data statistics (Apple Differential Privacy Team, 2017), and Microsoft's integration into Windows telemetry for application usage tracking (Ding et al., 2017).

A well-known drawback of LDP is the severe degradation of data representation utility due to its heavy randomization (Duchi et al., 2013; Kairouz et al., 2016). To provide rigorous privacy guarantees, standard LDP mechanisms enforce a pessimistic randomization process. Specifically, since the privacy guarantee must hold for any arbitrary points over the entire private data space, the magnitude of randomness is governed by the worst-case sensitivity. Although this prevents inferring any individual's true input, it significantly lowers the signal-to-noise ratio and distorts the original characteristics of the data. As a result, downstream tasks that require fine-grained feature preservation, such as nonlinear multi-class classification in high-dimensional spaces, remain challenging under the standard LDP paradigm (Duchi et al., 2018; Duan et al., 2022).

To mitigate this privacy–utility trade-off, a substantial body of literature has explored various directions. For example, prior work has focused on optimizing privacy mechanisms (Duchi et al., 2013; 2018; Wang et al., 2019; Bhowmick et al., 2018; Asi et al., 2022). To allocate privacy budgets and noise more effectively, researchers have also proposed task-aware (Cheng et al., 2022), relevance-aware (Phan et al., 2017), coordinate-wise (Alaggan et al., 2016; Muthukrishnan and Kalyani, 2025), and correlation-aware approaches (Kifer and Machanavajjhala, 2014; Aumüller et al., 2024; Dagan

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2024). Finally, another line of work has investigated post-processing methods (Erlingsson et al., 2014; Apple Differential Privacy Team, 2017; Ding et al., 2017; Hay et al., 2010; Wang et al., 2017; Jia and Gong, 2019; Cormode et al., 2019; Wang et al., 2020; Sajadmanesh and Gatica-Perez, 2021; Fang et al., 2023), as well as the incorporation of public data, models, and prior knowledge (Kurakin et al., 2022; Nasr et al., 2023; Hou et al., 2024), to improve utility while preserving privacy.

However, enhancing utility through the analysis of downstream models remains underexplored. Although Cheng et al. (2022) investigated this direction, their methodology is predominantly centered around linear models and requires the direct use of private data. Moreover, it does not fully utilize geometric principles. For example, the data representation space can be partitioned into a task-relevant row space and a task-irrelevant null space. These subspaces, which are distinct from the subspace reflecting the primary data correlations (Aumüller et al., 2024), dictate the directions in which the model is sensitive or robust to perturbations. While identifying and exploiting these subspaces can enhance utility, the existing LDP literature has largely overlooked these structural properties.

In this paper, we propose a novel approach that enhances data representation utility under LDP by reducing the amount of noise injected into task-critical subspaces defined by the downstream models. Our approach is motivated by the geometric insight that Jacobian-based analysis can identify task-sensitive and task-insensitive subspaces, thereby enabling fine-grained control over the noise injected into each subspace. Furthermore, we exploit the observation that sensitivity modulation can effectively reduce the noise scale applied to selected subspaces. Building on these principles, our approach transforms the isotropic noise introduced by the base randomizer into anisotropic noise, reducing its magnitude within the downstream model's task-sensitive subspace.

To the best of our knowledge, this work is the first to investigate improving representation utility under LDP through Jacobian-based identification of task-relevant subspaces and anisotropic noise reshaping. The main advantages of our approach are as follows: (i) it reduces the noise injected into task-critical subspaces, thereby substantially improving representation utility; (ii) it enables mechanisms satisfying pure LDP to outperform approximate LDP mechanisms in terms of utility; (iii) it is applicable to both linear and nonlinear downstream models; (iv) it can be seamlessly integrated with advanced privacy mechanisms by operating as a pre- and post-processing wrapper around a base mechanism; (v) it preserves the privacy guarantee of the base mechanism without incurring any additional privacy cost, as its transformations are determined solely by public information; and (vi) it enables effective nonlinear multi-class classification on high-dimensional data representations protected by LDP, while prior studies primarily focus on simpler tasks, such as frequency estimation, heavy hitter identification, mean estimation, and linear transformations.

2 Related Work

Task-Aware LDP. The work by Cheng et al. (2022) is most closely related to ours in that it utilizes task model information and serves as a wrapper compatible with other mechanisms: they propose an encoding-decoding framework that exploits downstream model information to apply LDP to high-dimensional data, injecting noise between the encoding and decoding steps. The most significant difference emerges when dealing with nonlinear downstream models. For these models, they propose training the encoder and decoder using unprotected private data. The encoder and decoder are trained by minimizing the discrepancy between the downstream model's outputs when fed with protected versus unprotected private data. In contrast, we utilize the Jacobian of the downstream model to isolate task-relevant from task-irrelevant subspaces. We also eliminate the risk of privacy leakage while improving performance as our subspace identification relies entirely on public information.

Coordinate-Wise DP. Our work is also aligned with (Muthukrishnan and Kalyani, 2025), which investigates the determination of non-identical noise scales for independent coordinates with varying sensitivities under DP. Their primary objective is to minimize the MSE between the original and privatized data. In contrast, our method prioritizes downstream utility over data MSE (DMSE) minimization. Specifically, our objective is to minimize a first-order surrogate for the risk caused by noise injection. These two objectives are often misaligned; for instance, injecting substantial noise into the null space may significantly degrade the DMSE while leaving downstream performance unaffected. Another key difference is that their method does not account for the task geometry. Finally, while they directly inject anisotropic noise into the private data, we inject isotropic noise, which is then distributed anisotropically through reshaping.

Public Data- and Model-Assisted DP. Public data and pre-trained models have been used to mitigate utility loss under both central DP and LDP. For instance, Nasr et al. (2023) use public data both to generate synthetic training data and to re-center gradient clipping, thereby improving the quality of noisy gradients in DP-SGD. Hou et al. (2024) use public LLMs to expand DP synthetic text. Under LDP, Jia and Gong (2019) incorporate public knowledge about the noise and the item distribution into frequency estimation, and Sajadmanesh and Gatica-Perez (2021) use the public graph topology to denoise private node features when training graph neural networks. Unlike prior work, we propose to utilize the Jacobian of a pre-trained model computed on public data to improve utility under LDP.

We discuss task/correlation-aware, feature-wise, and post-processing methods in Appendix A.1.

3 PRELIMINARIES

3.1 JACOBIAN-GUIDED TASK-SENSITIVE SUBSPACE IDENTIFICATION

Suppose that we have an m-dimensional data representation $\boldsymbol{z} \in \mathbb{R}^m$ and a linear task function $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^k$, which applies a weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times m}$ to a representation $\boldsymbol{z}$, i.e., $\mathcal{T}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$, where $m, k \in \mathbb{Z}^+$. Algebraically, the task-sensitive row space and the task-insensitive null space of $\boldsymbol{W}$ form an orthogonal decomposition of $\mathbb{R}^m$. The row space, $\text{Row}(\boldsymbol{W})$, is spanned by the row vectors of $\boldsymbol{W}$ and captures the directions along which the transformation has a non-trivial effect. An orthonormal basis for $\text{Row}(\boldsymbol{W})$, $\boldsymbol{u}_r$, is derived from the Jacobian, $\boldsymbol{J}_{\mathcal{T}}(\boldsymbol{z}) = \partial\mathcal{T}(\boldsymbol{z})/\partial\boldsymbol{z} = \boldsymbol{W}$, which is constant and equals the weight matrix. In contrast, the null space, $\text{Null}(\boldsymbol{W}) = \{\boldsymbol{z} \in \mathbb{R}^m \mid \boldsymbol{W}\boldsymbol{z} = \boldsymbol{0}\}$, consists of all vectors annihilated by the transformation. A basis for $\text{Null}(\boldsymbol{W})$, $\boldsymbol{u}_n$, can be found by computing an orthonormal set of vectors that span the orthogonal complement of the row space. These subspaces characterize which components of a representation are task-relevant and task-irrelevant, respectively.

For a nonlinear task function, unlike the linear case, a single global weight matrix that fully characterizes the transformation does not exist. Instead, we can analyze the local behavior of $\mathcal{T}$ around a specific representation $\boldsymbol{z}_0$ using a first-order Taylor expansion: $\mathcal{T}(\boldsymbol{z}_0 + \Delta\boldsymbol{z}) \approx \mathcal{T}(\boldsymbol{z}_0) + \boldsymbol{J}_{\mathcal{T}}(\boldsymbol{z}_0)\Delta\boldsymbol{z}$, where $\Delta\boldsymbol{z}$ is a sufficiently small perturbation. Locally, the Jacobian $\boldsymbol{J}_{\mathcal{T}}(\boldsymbol{z}_0)$ acts analogously to the linear weight matrix $\boldsymbol{W}$, allowing us to define a local row space and a local null space specific to the point $\boldsymbol{z}_0$. If the perturbation $\Delta\boldsymbol{z}$ lies within the local null space of $\boldsymbol{J}_{\mathcal{T}}(\boldsymbol{z}_0)$, then $\boldsymbol{J}_{\mathcal{T}}(\boldsymbol{z}_0)\Delta\boldsymbol{z} = \boldsymbol{0}$, and consequently $\mathcal{T}(\boldsymbol{z}_0 + \Delta\boldsymbol{z}) \approx \mathcal{T}(\boldsymbol{z}_0)$, meaning that the nonlinear transformation remains locally invariant to perturbations along the null space directions.

3.2 LOCAL DIFFERENTIAL PRIVACY

Local Differential Privacy (LDP) (Kasiviswanathan et al., 2011) applies a randomization mechanism to each data representation $\boldsymbol{z} \in \mathcal{Z}$, within the representation space $\mathcal{Z} \subset \mathbb{R}^m$. Let $\epsilon > 0$ denote the privacy budget and $\delta \geq 0$ the failure probability, where $\delta = 0$ corresponds to pure ϵ-LDP and $\delta > 0$ to (ϵ, δ)-LDP. A mechanism $\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^m$ satisfies (ϵ, δ)-LDP if, for any pair of representations $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$ and any measurable subset $\mathcal{S} \subseteq \text{Range}(\mathcal{M})$, it holds that $\Pr[\mathcal{M}(\boldsymbol{z}) \in \mathcal{S}] \leq e^{\epsilon} \Pr[\mathcal{M}(\boldsymbol{z}') \in \mathcal{S}] + \delta$. A canonical class of mechanisms satisfying these guarantees is that of additive noise mechanisms (Dwork and Roth, 2014) of the form (1):

$$\tilde{\boldsymbol{z}} = \mathcal{M}(\boldsymbol{z}) = \boldsymbol{z} + \boldsymbol{\xi}. \tag{1}$$

For $(\epsilon, 0)$-LDP (i.e., pure ϵ-LDP), the Laplace mechanism $\mathcal{M}_L$ is a natural choice, where the noise scale b is directly proportional to the ℓ_1-sensitivity, defined as $\Delta_1 = \max_{\boldsymbol{z},\boldsymbol{z}' \in \mathcal{Z}} \|\boldsymbol{z} - \boldsymbol{z}'\|_1$, and each element of $\boldsymbol{\xi}$ is sampled i.i.d. from $\text{Lap}(0, \Delta_1/\epsilon)$ (Dwork and Roth, 2014). For (ϵ, δ)-LDP with $\delta > 0$, the Gaussian mechanism $\mathcal{M}_G$ is a natural choice, where $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\mathbf{I})$ and the noise scale σ is proportional to the ℓ_2-sensitivity, $\Delta_2 = \max_{\boldsymbol{z},\boldsymbol{z}' \in \mathcal{Z}} \|\boldsymbol{z} - \boldsymbol{z}'\|_2$, calibrated via the Analytic Gaussian Mechanism (Balle and Wang, 2018). As a local instantiation of DP, LDP inherits its fundamental strengths, including worst-case bounds on privacy leakage and immunity to post-processing (Dwork and Roth, 2014). These theoretical bounds hold even against an adversary with full knowledge of the randomization mechanism $\mathcal{M}$ and its parameters, including the privacy budget, sensitivity, and noise distribution settings. The privacy guarantees of LDP are preserved under arbitrary post-processing independent of (private) data representation. Once $\tilde{\boldsymbol{z}}$ is produced by a mechanism satisfying (ϵ, δ)-LDP, any measurable downstream computation or transformation $g(\tilde{\boldsymbol{z}})$ independent of the private data maintains the same privacy guarantee without consuming additional privacy budget.

4 MOTIVATION

Our primary objective is to enhance the utility of multi-dimensional representations randomized by an LDP mechanism, where sensitivities scale with dimensionality (i.e., $\Delta_1 = \mathcal{O}(m)$ and $\Delta_2 = \mathcal{O}(\sqrt{m})$(Mireshghallah et al., 2020; 2021; Cheng et al., 2022; Mai et al., 2024)). To achieve this, we explore a novel approach that mitigates the noise injected into task-relevant subspaces. Although we primarily focus on pure ϵ-LDP, we seek an approach that seamlessly extends to (ϵ, δ)-LDP ($\delta > 0$).

Our approach is motivated by the observation that a classifier exhibits anisotropic sensitivity to perturbations in the representation space. Consider a linear binary classifier, $\mathcal{T}(\boldsymbol{z}) = \boldsymbol{w}^\top \boldsymbol{z} + b$, with a decision boundary in $\mathbb{R}^2$. The decision boundary is defined by a normal vector $\boldsymbol{w} \in \mathbb{R}^2$, and the space can be spanned by two basis vectors $\boldsymbol{u}_r$ and $\boldsymbol{u}_n$, aligned parallel and perpendicular to $\boldsymbol{w}$, respectively. $\boldsymbol{u}_r$ and $\boldsymbol{u}_n$ span the row space and null space of the classifier, respectively. Crucially, perturbing a representation $\boldsymbol{z}$ along $\boldsymbol{u}_r$ can alter the classification outcome, whereas perturbations of any arbitrary magnitude along $\boldsymbol{u}_n$ leave the classification result unchanged. We provide an illustration of this motivating example in Appendix A.2. Based on this observation, our approach anisotropically reshapes the isotropic noise $\boldsymbol{\xi}$ to allocate less noise along the task-sensitive direction $\boldsymbol{u}_r$ and more along the task-invariant direction $\boldsymbol{u}_n$.

5 PROPOSED APPROACH

We propose to reduce the noise injected into the task-critical subspaces defined by the server-side downstream model. Our approach comprises a pre-processing function $f : \mathcal{Z} \to \bar{\mathcal{Z}}$, where $\bar{\mathcal{Z}}$ is an intermediate bounded space in $\mathbb{R}^m$, and a post-processing function $g : \mathbb{R}^m \to \mathbb{R}^m$. A randomization mechanism $\mathcal{M}$ is applied to the intermediate representation $\bar{\boldsymbol{z}} = f(\boldsymbol{z})$, followed by the application of the post-processing function g. As a result, this process yields a randomized version of $\boldsymbol{z}$ injected with anisotropic noise $\boldsymbol{\xi}_a \in \mathbb{R}^m$ in (2):

$$\hat{\boldsymbol{z}} = g(\mathcal{M}(f(\boldsymbol{z}))) = g(\mathcal{M}(\bar{\boldsymbol{z}})) = g(\bar{\boldsymbol{z}} + \boldsymbol{\xi}) \approx \boldsymbol{z} + \boldsymbol{\xi}_a, \tag{2}$$

where the covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times m}$ of $\boldsymbol{\xi}_a$ is publicly known and positive-definite. This matrix is designed to rotate and scale the noise, attenuating it along the row space of the downstream model, while amplifying it along the null space. A more detailed description is provided in Appendix A.3.

5.1 PRE-PROCESSING AND POST-PROCESSING

Suppose that the eigendecomposition of $\boldsymbol{\Sigma}$ is given by $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is an orthogonal matrix representing rotation, and $\boldsymbol{\Lambda} \in \mathbb{R}^{m \times m}$ is a diagonal matrix that scales each coordinate in $\mathbb{R}^m$. We derive our pre-processing and post-processing functions by modifying (3):

$$\boldsymbol{z} + \boldsymbol{\xi}_a = \boldsymbol{L}(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}) + \boldsymbol{\xi}) + \boldsymbol{\mu} \tag{3}$$

where $\boldsymbol{\mu} \in \mathbb{R}^m$ represents an offset and $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2} \in \mathbb{R}^{m \times m}$ denotes a reshaping factor.

Pre-processing. Define the configuration tuple as $\boldsymbol{\phi} = (\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$, and let $\rho > 0$ and $p \in [1, \infty]$. Incorporating clipping, we define our pre-processing function $f_{\boldsymbol{\phi}} : \mathcal{Z} \to \bar{\mathcal{Z}}$ as (4):

$$f_{\boldsymbol{\phi}}(\boldsymbol{z}; p, \rho) = \theta_p\left(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}); \rho\right) \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}) \tag{4}$$

where $\theta_p(\cdot; \rho)$ denotes ℓ_p-norm-based radial clipping with a threshold ρ. The detailed construction of $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ via Jacobian computation is presented in Section 5.2. Using $\boldsymbol{L}^{-1}$, this pre-processing maps the representation $\boldsymbol{z}$ from the space $\mathcal{Z} \subset \mathbb{R}^m$ to another bounded space $\bar{\mathcal{Z}} \subset \mathbb{R}^m$. Specifically, it first offsets $\boldsymbol{z}$ by $\boldsymbol{\mu}$, applies an inverse rotation and rescaling via $\boldsymbol{L}^{-1}$, and subsequently bounds the result to ensure that $\bar{\boldsymbol{z}} \in \bar{\mathcal{Z}}$ strictly resides within a bounded space.

Post-processing. We define the post-processing function $g_{\boldsymbol{\phi}} : \mathbb{R}^m \to \mathbb{R}^m$ as (5):

$$g_{\boldsymbol{\phi}}(\mathcal{M}(f_{\boldsymbol{\phi}}(\boldsymbol{z}; p, \rho))) = \boldsymbol{L}(f_{\boldsymbol{\phi}}(\boldsymbol{z}; p, \rho) + \boldsymbol{\xi}) + \boldsymbol{\mu} \approx \boldsymbol{z} + \boldsymbol{\xi}_a. \tag{5}$$

Our post-processing function g reverses the transformation $\boldsymbol{L}^{-1}$ and the offset $\boldsymbol{\mu}$ applied in f, while simultaneously rotating and rescaling the isotropic noise $\boldsymbol{\xi}$ injected into $\bar{\mathcal{Z}}$. This yields a new anisotropic noise vector $\boldsymbol{\xi}_a$ designed to enhance downstream utility. Notably, by virtue of post-processing immunity, this approach preserves the privacy guarantees because g is applied to the output of the ϵ-LDP mechanism $\mathcal{M}$.

5.2 Covariance Matrix for Rotation and Rescaling

The construction of $\boldsymbol{\Sigma}$ consists of two components: the *rotation* and the *scaling*, as mentioned in Section 5.1. We first determine the rotation matrix $\boldsymbol{U}$, which aligns the axes of $\boldsymbol{\xi}_a$ with the basis vectors of the task-sensitive and task-insensitive subspaces. Then, we determine the scaling matrix $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \cdots, \lambda_m) \succ 0$, which assigns noise scales to each coordinate to enhance data utility.

Rotation. To identify the relevant basis vectors for rotation, we characterize the global geometric behavior of the downstream models. First, we evaluate the local Jacobian matrices $\boldsymbol{J}_i \in \mathbb{R}^{k\times m}$ for $i \in \{1, \ldots, N\}$ at N samples from public datasets. Next, we construct an aggregated Jacobian matrix $\boldsymbol{J} = \frac{1}{\sqrt{N}} \begin{bmatrix} \boldsymbol{J}_1^\top & \cdots & \boldsymbol{J}_N^\top \end{bmatrix}^\top \in \mathbb{R}^{Nk\times m}$ by vertically stacking $\boldsymbol{J}_i$. Subsequently, we obtain the Gram matrix of this empirically aggregated matrix, $\widehat{\boldsymbol{\Omega}} = \boldsymbol{J}^\top \boldsymbol{J}$, and perform eigendecomposition on $\widehat{\boldsymbol{\Omega}}$ such that $\widehat{\boldsymbol{\Omega}} = \boldsymbol{V}\,\mathrm{diag}(\gamma_1, \ldots, \gamma_m)\boldsymbol{V}^\top$, where $\mathrm{diag}(\gamma_1, \ldots, \gamma_m) \in \mathbb{R}^{m\times m}$ is a positive semi-definite diagonal matrix containing the eigenvalues, and $\boldsymbol{V} \in \mathbb{R}^{m\times m}$ represents the eigenvectors of $\widehat{\boldsymbol{\Omega}}$ used for our rotation ($\boldsymbol{U} = \boldsymbol{V}$). The eigenvectors in $\boldsymbol{U}$ associated with significant eigenvalues span the global active subspace (Constantine et al., 2014), representing the task-sensitive directions. Conversely, the vectors corresponding to zero eigenvalues span the (task-insensitive) inactive subspace.

Scaling. The objective of scaling is to assign a tailored noise scale factor σ_i to each coordinate $i \in \{1, 2, \ldots, m\}$ of the noise $\boldsymbol{\xi}_a = \boldsymbol{L}\boldsymbol{\xi}$, where these coordinates are aligned with the basis vectors of the task-sensitive and task-insensitive subspaces. To determine these individual scale factors, we use the eigenvalues γ_i of the Jacobian Gram matrix, as they represent the relative importance of the i-th coordinate. Additionally, we impose $p = 1$ and a gauge constraint such that the sum of the inverse noise scales remains constant before and after rescaling; that is, $m = \sum_{i=1}^{m} (\lambda_i)^{-1/2}$, where $\lambda_i = (\sigma_i/\sigma)^2$ and $0 < \sigma_i < \infty$. The isotropic noise scale σ is determined once ρ is set. Furthermore, to prevent the risk from diverging due to the empirical Jacobian, the distribution discrepancy between public and private data, and the nonlinear remainder, we assign a finite scale cap $1 < \lambda_{\max} < \infty$ to the task-insensitive coordinates associated with zero or very small eigenvalues. Let C and F denote the sets of capped and free coordinates, respectively. The Lagrange multipliers yields scales of (6):

$$\lambda_i^\star = \begin{cases} \lambda_{\max}, & i \in C, \\ \left(\dfrac{m - |C|\lambda_{\max}^{-1/2}}{\sum_{j\in F} \gamma_j^{1/3}} \gamma_i^{1/3} \right)^{-2}, & i \in F. \end{cases} \tag{6}$$

This is obtained by jointly optimizing the Jacobian basis and the scale allocation, using the noise component of the first-order risk surrogate as the objective function. This choice of objective is inspired by our observation that the noise scale, rather than the distortion induced by clipping, is the primary bottleneck under LDP constraints (see Appendix A.10.7 for details). A detailed analysis of the derivation, joint optimality of the Jacobian basis and scale allocation (Propositions 1 and 2), performance bounds (Propositions 3, 4 and 5), and prediction stability (Proposition 6) is provided in Appendix A.4.

5.3 Privacy Guarantee

Our configuration parameters $\boldsymbol{U}$, $\boldsymbol{\Lambda}$, $\boldsymbol{\mu}$, p, and ρ are selected using public information and fixed before privatization. Privacy then follows from the bounded-domain property of $f_{\boldsymbol{\phi}}$, the privacy guarantee of the base randomizer $\mathcal{M}$ on that domain, and post-processing by $g_{\boldsymbol{\phi}}$.

Theorem 1 (Privacy Guarantee of Jacobian-Guided Anisotropic Noise Reshaping)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^m$ be the domain of a single private representation, which may possibly be unbounded. Assume:*

- ***(A1) Public configuration.*** *Conditional on fixed public side information, let $\boldsymbol{U}$ be orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ satisfy $0 < \lambda_{\min} \le \lambda_i \le \lambda_{\max} < \infty$. Define the configuration tuple $\boldsymbol{\phi} = (\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$, and let $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ (which is invertible and has a finite condition number). Let $\rho > 0$ and $p \in [1, \infty]$. $\boldsymbol{\phi}$, p, and ρ may be chosen analytically or empirically using only the public task model, public data, and public validation results. Once selected, they are fixed before any input is privatized, and are neither computed from nor adapted to the protected representation.*

- ***(A2) Bounded intermediate domain.*** *With $\theta_p(\bar{\boldsymbol{z}}; \rho) = \min(1, \rho/\|\bar{\boldsymbol{z}}\|_p)$ for $\bar{\boldsymbol{z}} \neq \mathbf{0}$ and 1 otherwise, define $f_\phi(\boldsymbol{z}) = \theta_p\big(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}); \rho\big)\, \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$ and the bounded domain $\bar{\mathcal{Z}}_{p,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_p \leq \rho\}$. Radial clipping ensures $f_{\boldsymbol{\phi}}(\mathcal{Z}) \subseteq \bar{\mathcal{Z}}_{p,\rho}$ for every $\mathcal{Z}$, regardless of whether it is bounded or not.*
- ***(A3) Underlying randomizer.*** *The base randomizer $\mathcal{M}$ is instantiated with the domain and calibration required by its original privacy theorem. Under the corresponding conditions, $\mathcal{M}$ satisfies (ϵ, δ)-LDP uniformly over $\bar{\mathcal{Z}}_{p,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_p \leq \rho\}$.*
- ***(A4) Post-processing.*** *$g_{\boldsymbol{\phi}}(\boldsymbol{v}) = \boldsymbol{L}\boldsymbol{v} + \boldsymbol{\mu}$ is deterministic, measurable, and fixed solely from public information.*

Then for every admissible configuration $\boldsymbol{\phi}$, the complete mechanism $\Phi_{\boldsymbol{\phi}} = g_{\boldsymbol{\phi}} \circ \mathcal{M} \circ f_{\boldsymbol{\phi}}$ satisfies (ϵ, δ)-LDP on $\mathcal{Z}$.

We provide the proof, detailed instantiations of mechanisms, including the Laplace, Analytical Gaussian Mechanism, and PrivUnit, and additional properties in Appendices A.6–A.8.

6 Evaluation

We evaluate the utility of our method by measuring predictive performance across downstream inference tasks (one regression and two image classification tasks) under varying representation dimensions ($m \in \{16, 32, 64\}$) and privacy budgets ($\epsilon \in [0.5, 10]$).

Datasets. We apply user-level LDP to time-series regression using London household smart meter (LHSM) data (UK Power Networks; D., 2019), which records the half-hourly electricity consumption (kWh/half-hour) of 5,547 London households over approximately 27 months. Additionally, we apply item-level LDP (see Corollary 8) to image classification on the CIFAR-10-C, CIFAR-10 datasets Krizhevsky (2009), and MNIST (Lecun et al., 1998).

We split the LHSM dataset into public and private sets based on both household and chronological criteria. The public dataset comprises the initial 14 months of data from 3,000 randomly sampled households, used for training the LR model. The private dataset consists of the remaining 13 months from the other households, where it is independent and out-of-sample for the pre-trained model.

The training sets from CIFAR-10 and MNIST are treated as public. These are partitioned into training and validation sets with an 80:20 split to pre-train the public feature extractors (VAE (Kingma and Welling, 2014) and ResNet-20 (He et al., 2016)) and classifiers. To assess the utility of the mechanisms, the test sets of CIFAR-10-C Hendrycks and Dietterich (2019) (e.g., brightness, fog, defocus blur, and Gaussian noise), CIFAR-10, and MNIST are treated as private. Note that the CIFAR-10-C test set exhibits a distribution shift compared to the CIFAR-10 training set.

Downstream Tasks. Both the regression and classification models are trained on the public dataset and subsequently evaluated by feeding them the randomized private dataset to measure accuracy and agreement. Specifically, the regression task involves predicting the next day's *energy_mean* using the previous 16 days ($m = 16$) of *energy_mean* values as input. For the classification task, an m-dimensional feature representation ($m \in \{32, 64\}$) is first extracted using a public feature extractor and then fed into the classifier.

Downstream Models. We employ linear regression (LR), linear classifiers (LC), and multi-layer perceptron (MLP) classifiers with ReLU, GELU, and Tanh activations.

Baselines. Our evaluation includes foundational mechanisms, such as the standard Laplace (Dwork and Roth, 2014), AGM (Balle and Wang, 2018), and PrivUnit2 (Bhowmick et al., 2018), and state-of-the-art approaches, including Instance Optimal (IO) (Huang et al., 2021), PrivUnitG (Asi et al., 2022), Task-Aware (Cheng et al., 2022), PLAN (Aumüller et al., 2024), and Coordinate-wise (CW) (Muthukrishnan and Kalyani, 2025).

6.1 Results: Privacy-Utility Trade-offs

We evaluate utility using Root Mean Squared Error (RMSE) for regression and average accuracy for classification, where lower RMSE and higher accuracy indicate better performance, respectively.

Table 1: Main results for **Linear Regression** ($m = 16$) on the **LHSM**. The Root Mean Square Error (RMSE) is used. PA denotes the proposed approach.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691
Laplace (ℓ_1)	17.6293	8.8154	5.8778	4.4093	3.5284	2.9413	2.5221	2.2078	1.9635	1.7682	1.1832	0.8921
Laplace+PA	**1.0520**	**0.5387**	**0.3729**	**0.2935**	**0.2482**	**0.2197**	**0.2006**	**0.1872**	**0.1774**	**0.1700**	**0.1511**	**0.1439**
PrivUnit2	3.9709	2.0035	1.3492	1.0252	0.8348	0.7097	0.6222	0.5577	0.5083	0.4710	0.3619	0.3155
PrivUnit2+PA	**0.9048**	**0.4776**	**0.3437**	**0.2822**	**0.2490**	**0.2282**	**0.2150**	**0.2058**	**0.1991**	**0.1945**	**0.1826**	**0.1780**
PrivUnitG	3.9645	1.9924	1.3471	1.0290	0.8425	0.7192	0.6333	0.5707	0.5235	0.4865	0.3831	0.3365
PrivUnitG+PA	**0.8996**	**0.4736**	**0.3406**	**0.2801**	**0.2471**	**0.2269**	**0.2136**	**0.2044**	**0.1981**	**0.1936**	**0.1822**	**0.1765**
CW+PA w/o bounding	**1.1897**	**0.5976**	**0.4016**	**0.3045**	**0.2470**	**0.2093**	**0.1828**	**0.1634**	**0.1486**	**0.1371**	**0.1048**	**0.0908**
Task-Aware	0.1920	0.1920	0.1918	0.1916	0.1913	0.1910	0.1906	0.1901	0.1896	0.1890	0.1852	0.1802
AGM ($\ell_2, \delta = 10^{-5}$)	11.6732	6.1941	4.2890	3.3123	2.7156	2.3121	2.0205	1.7995	1.6261	1.4863	1.0593	0.8400
IO ($\delta = 10^{-5}$)	8.2244	5.9637	4.8461	4.1808	3.7330	3.3761	3.0977	2.8843	2.7001	2.5527	2.0428	1.7520
PLAN ($\delta = 10^{-5}$)	8.4207	4.2118	2.8095	2.1088	1.6888	1.4091	1.2095	1.0601	0.9441	0.8514	0.5753	0.4395

Table 2: Main results for the **Nonlinear MLP Classifier with ReLU** ($m = 64$) on the **CIFAR-10-C (Brightness, Severity 5)**. Average accuracy is used. PA denotes the proposed approach.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365
Laplace (ℓ_1)	0.1001	0.1013	0.1025	0.1038	0.1050	0.1063	0.1074	0.1087	0.1099	0.1111	0.1177	0.1251
Laplace+PA	**0.1048**	**0.1103**	**0.1161**	**0.1219**	**0.1280**	**0.1342**	**0.1405**	**0.1467**	**0.1538**	**0.1609**	**0.1989**	**0.2416**
PrivUnit2	0.1099	0.1203	0.1316	0.1434	0.1560	0.1686	0.1820	0.1950	0.2092	0.2227	0.2933	0.3615
PrivUnit2+PA	**0.1194**	**0.1410**	**0.1660**	**0.1936**	**0.2223**	**0.2536**	**0.2840**	**0.3152**	**0.3453**	**0.3754**	**0.4978**	**0.5840**
PrivUnitG	0.1083	0.1190	0.1299	0.1415	0.1541	0.1665	0.1798	0.1931	0.2061	0.2197	0.2858	0.3375
PrivUnitG+PA	**0.1201**	**0.1425**	**0.1672**	**0.1936**	**0.2227**	**0.2524**	**0.2822**	**0.3121**	**0.3414**	**0.3688**	**0.4860**	**0.5648**
CW+PA w/o bounding	**0.1025**	**0.1058**	**0.1092**	**0.1128**	**0.1164**	**0.1201**	**0.1241**	**0.1282**	**0.1323**	**0.1366**	**0.1581**	**0.1828**
AGM ($\ell_2, \delta = 10^{-5}$)	0.1023	0.1046	0.1068	0.1090	0.1116	0.1139	0.1157	0.1181	0.1202	0.1225	0.1321	0.1423
IO ($\delta = 10^{-5}$)	0.1045	0.1080	0.1112	0.1142	0.1171	0.1198	0.1226	0.1253	0.1280	0.1307	0.1433	0.1561
PLAN ($\delta = 10^{-5}$)	0.1022	0.1049	0.1080	0.1107	0.1134	0.1163	0.1194	0.1224	0.1260	0.1290	0.1464	0.1654

Tables 1 and 2 report the conservative results of the proposed approach (PA) over 20 different seeds: RMSE across 2,547 households for regression, and average accuracy over 10,000 samples for nonlinear classification using CIFAR-10-C (brightness). Further details on our experimental setup, evaluation protocol and dataset descriptions are provided in Appendix A.10. We here present the primary results for downstream models, specifically LR and an MLP classifier with ReLU activations.

Linear Regression. Mechanisms without PA exhibited high RMSE values under low privacy budgets. At $\epsilon = 0.5$, Laplace recorded 17.63 kWh/half-hour, while PrivUnit2 and PrivUnitG reached 3.97 and 3.96 kWh/half-hour, respectively. Although RMSE decreased as ϵ increased, even at $\epsilon = 10.0$, Laplace remained at 0.89 kWh/half-hour and PrivUnit2 at 0.32 kWh/half-hour, indicating limited utility without PA. Task-Aware approach maintained its performance largely regardless of the budget; yet, in tasks that deviate from their MSE optimization objective, such as classification, we observed collapse, with performance falling below that of Laplace (Table 11 in Appendix A.10.9).

Mechanisms augmented with PA achieved lower RMSE across all ϵ values. Laplace+PA recorded 1.05 kWh/half-hour at $\epsilon = 0.5$ and 0.54 kWh/half-hour at $\epsilon = 1.0$, reducing RMSE by factors of approximately 16.8 and 16.4, respectively, compared with the base Laplace mechanism. PrivUnit2+PA and PrivUnitG+PA demonstrated comparable improvements, and they outperformed Laplace+PA. CW+PA w/o bounding also exhibited competitive performance. This may show that CW's DMSE optimization translates to linear problems computed in the form of $\boldsymbol{W}\boldsymbol{z}$. It recorded 0.14 kWh/half-hour at $\epsilon = 5.0$ and 0.10 kWh/half-hour at $\epsilon = 7.5$, outperforming all other mechanisms and approaching the no-randomization baseline of 0.07 kWh/half-hour most closely. While approximate mechanisms, AGM, IO, and PLAN, outperform the standard Laplace mechanism, Laplace+PA surpasses them all. These approximate mechanisms also fall short of CW+PA w/o bounding. Considering that

approximate mechanisms are specifically designed to improve utility by tolerating a privacy failure probability, it is significant that PA-integrated approaches substantially outperform them.

Nonlinear Classification. The nonlinear classification experiments were conducted on the CIFAR-10-C (Brightness corruption at the highest severity level 5) test dataset using a ResNet-20-based nonlinear MLP classifier with ReLU activations, where 64-dimensional feature vectors serve as input. Utility is evaluated by classification accuracy, with the no-randomization baseline achieving 0.8365. Baseline mechanisms without PA exhibited severely degraded performance across all privacy budgets. Laplace (ℓ_1) yielded accuracies near 0.1, equivalent to random guessing on a 10-class task, while PrivUnit2 and PrivUnitG achieved results of 0.2933 and 0.2858 at $\epsilon = 7.5$, respectively.

The task-aware approach by Cheng et al. (2022) is focused on MNIST and does not offer guidance for more complex datasets such as CIFAR-10(-C). Crucially, their method relies on direct access to private data during the training phase. Despite avoiding such direct access to private data, our approach demonstrates superior results. As detailed in Appendix A.10.6, integrating PA with PrivUnit2 and PrivUnitG significantly outperforms their method under its optimal configuration, while the Laplace+PA mechanism yields comparable or superior performance.

The integration of PA consistently and significantly improved classification accuracy across all mechanisms and ϵ values. Laplace+PA outperformed its base counterpart at every operating point, achieving an accuracy of 0.1989 at $\epsilon = 7.5$, compared to 0.1177 without PA. Furthermore, Laplace+PA surpassed all evaluated approximate mechanisms. More notably, PrivUnit2+PA and PrivUnitG+PA delivered substantially stronger performance, achieving accuracies of 0.4978 and 0.4860 at $\epsilon = 7.5$ and 0.3754 and 0.3688 at $\epsilon = 5.0$, respectively. This represents an improvement of approximately 20 and 15 percentage points for each respective budget compared to their non-PA counterparts, underscoring the effectiveness of PA in recovering a significant portion of the utility lost to privacy noise. CW+PA w/o bounding, however, underperformed relative to other PA-augmented mechanisms in this setting, recording only 0.1581 at $\epsilon = 7.5$ and 0.1366 at $\epsilon = 5.0$. This performance was even lower than that of Laplace+PA. The distinction between these two approaches lies in the use of sensitivity bounding versus a DMSE-optimized noise scale. These results suggest that our clipping-based approach is more effective under conditions where DMSE optimization is suboptimal.

6.2 Ablation Study

Pre-processing and Post-processing. We conduct an ablation study to investigate the individual contributions of the pre-processing and post-processing components within PA.

In summary, the ablation results demonstrate that both pre-processing and post-processing components are indispensable. Table 3 provides the results across all three mechanisms on CIFAR-10-C under brightness corruption at severity level 5. Removing either component causes all three mechanisms to collapse to near-random performance (10%), regardless of the privacy budget. A notable instability is observed when post-processing is ablated: the standard deviation of accuracy grows substantially with increasing ϵ, reaching up to 0.026 for PrivUnit2+PA and 0.025 for PrivUnitG+PA at $\epsilon = 10$. This suggests that without post-processing, the representations become erratic as the privacy budget relaxes, undermining the reliability of downstream classification.

To disentangle the internal mechanics of the post-processing phase, we evaluate a variant, which retains the noise rotation but omits the scaling procedure: $(\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}})$. Across Laplace, PrivUnit2, and PrivUnitG, the performance of this variant falls between the fully ablated post-processing baseline and the full PA integration. For instance, with PrivUnit2 at $\epsilon = 10$, the variant recovers substantial utility (0.5117) compared to the complete absence of post-processing (0.0875), yet still falls short of the full PA (0.5829). This result implies that while directional alignment (rotation) alone successfully mitigates a portion of the noise impact, structural scaling (reshaping) is necessary to fully utilize the proposed approach.

A further distinction emerges when comparing the variant without scaling against the direct ℓ_1-clipping baseline applied in the original space $\mathcal{Z}$. Although both configurations omit anisotropic noise scaling, they differ in one fundamental respect. The variant without scaling applies ℓ_1-clipping within the transformed space $\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$, where the axes are not only aligned with the task-critical subspaces but also inversely scaled according to their respective importance. In contrast, the direct ℓ_1-clipping baseline operates in the raw space $\mathcal{Z}$ without any such alignment. The consistent

Table 3: Ablation study on **Pre-Processing** and **Post-Processing** for **Laplace + PA**, **PrivUnit2+PA**, and **PrivUnitG+PA** on **CIFAR-10-C (Brightness, Severity 5)**. Average accuracy is used.

Mechanism	Privacy Budget (ϵ)											
	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace+PA ($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	**0.1045**	**0.1100**	**0.1161**	**0.1217**	**0.1279**	**0.1344**	**0.1411**	**0.1478**	**0.1543**	**0.1615**	**0.1996**	**0.2428**
Laplace+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}$)	0.1030	0.1048	0.1070	0.1087	0.1109	0.1130	0.1148	0.1171	0.1188	0.1210	0.1326	0.1447
Laplace+(ℓ_1)	0.1001	0.1015	0.1028	0.1042	0.1056	0.1068	0.1081	0.1096	0.1109	0.1124	0.1196	0.1276
Laplace+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{PCA}}^{-1}, \ell_1, \boldsymbol{U}_{\text{PCA}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.1014	0.1015	0.1016	0.1016	0.1017	0.1017	0.1019	0.1019	0.1019	0.1020	0.1022	0.1024
Laplace+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Rand}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Rand}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.1008	0.1008	0.1008	0.1008	0.1008	0.1008	0.1009	0.1009	0.1009	0.1009	0.1010	0.1011
Laplace+PA w/o Pre-processing	0.1015	0.1014	0.1014	0.1014	0.1009	0.1009	0.1010	0.1009	0.1007	0.1008	0.1006	0.1001
Laplace+PA w/o Post-processing	0.0984	0.0982	0.0979	0.0977	0.0976	0.0974	0.0971	0.0969	0.0966	0.0966	0.0954	0.0945
PrivUnit2+PA ($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	**0.1196**	**0.1410**	**0.1662**	**0.1945**	**0.2244**	**0.2553**	**0.2856**	**0.3171**	**0.3471**	**0.3760**	**0.4986**	**0.5829**
PrivUnit2+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}$)	0.1155	0.1319	0.1509	0.1707	0.1926	0.2148	0.2381	0.2614	0.2856	0.3104	0.4209	0.5117
PrivUnit2+(ℓ_1)	0.1111	0.1214	0.1328	0.1443	0.1567	0.1687	0.1820	0.1944	0.2092	0.2225	0.2942	0.3620
PrivUnit2+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{PCA}}^{-1}, \ell_1, \boldsymbol{U}_{\text{PCA}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.1000	0.1001	0.1002	0.1003	0.1003	0.1002	0.1005	0.1008	0.1011	0.1013	0.1015	0.1019
PrivUnit2+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Rand}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Rand}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.0993	0.0995	0.0997	0.0999	0.1001	0.1004	0.1007	0.1012	0.1019	0.1030	0.1043	0.1039
PrivUnit2+PA w/o Pre-processing	0.1000	0.0994	0.0999	0.0997	0.0997	0.0994	0.0990	0.0986	0.0986	0.0984	0.0976	0.0955
PrivUnit2+PA w/o Post-processing	0.0995	0.0987	0.0979	0.0971	0.0961	0.0954	0.0947	0.0940	0.0935	0.0931	0.0890	0.0875
PrivUnitG+PA ($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	**0.1205**	**0.1421**	**0.1666**	**0.1931**	**0.2230**	**0.2539**	**0.2838**	**0.3147**	**0.3440**	**0.3725**	**0.4887**	**0.5669**
PrivUnitG+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Jac}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Jac}}$)	0.1146	0.1312	0.1501	0.1700	0.1910	0.2142	0.2373	0.2607	0.2839	0.3078	0.4121	0.4869
PrivUnitG+(ℓ_1)	0.1089	0.1196	0.1303	0.1420	0.1551	0.1679	0.1812	0.1942	0.2071	0.2210	0.2864	0.3378
PrivUnitG+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{PCA}}^{-1}, \ell_1, \boldsymbol{U}_{\text{PCA}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.0992	0.0993	0.0994	0.0995	0.0995	0.0996	0.0998	0.0998	0.1000	0.1001	0.1005	0.1009
PrivUnitG+($\boldsymbol{\Lambda}_{\text{Jac}}^{-1/2}\boldsymbol{U}_{\text{Rand}}^{-1}, \ell_1, \boldsymbol{U}_{\text{Rand}}\boldsymbol{\Lambda}_{\text{Jac}}^{1/2}$)	0.1012	0.1014	0.1016	0.1018	0.1020	0.1022	0.1025	0.1027	0.1030	0.1032	0.1041	0.1049
PrivUnitG+PA w/o Pre-processing	0.0986	0.0987	0.0985	0.0979	0.0977	0.0979	0.0976	0.0975	0.0975	0.0978	0.0966	0.0958
PrivUnitG+PA w/o Post-processing	0.0986	0.0983	0.0973	0.0964	0.0957	0.0951	0.0940	0.0932	0.0924	0.0919	0.0889	0.0869

performance gap between these two variants across all three mechanisms demonstrates that clipping in a task-aligned space is inherently beneficial even without rescaling. This highlights that the utility of PA stems from the interaction of three key components: space alignment, sensitivity bounding in the transformed space, and anisotropic rescaling.

We also experimented with replacing the Jacobian-guided rotation $\boldsymbol{U}_{\text{Jac}}$ with PCA- or random permutation-based rotations: $\boldsymbol{U}_{\text{PCA}}$ and $\boldsymbol{U}_{\text{Rand}}$. These variants resulted in near-random performance, highlighting the necessity of the Jacobian-guided rotation. Furthermore, compared to the ℓ_1-clipping baseline, the full PA integration achieves up to 90% relative improvement (Laplace, $\epsilon = 10$), demonstrating that the utility gains stem not from clipping alone, but from the coordinated interaction of the pre- and post-processing pipeline introduced by PA.

6.3 Additional Experiments: Clipping Thresholds, Distribution Shift, Alternative Activations, and Aggregation

We present ablation studies on clipping thresholds, an analysis of distribution shifts across 20 corruption-severity combinations, experiments with alternative activation functions (e.g., GELU and Tanh), and aggregation results in Appendices A.10.7, A.10.8, A.10.9, and A.10.11, respectively. Our results indicate that (i) the noise scale, rather than clipping-induced distortion, is the primary bottleneck under LDP constraints; (ii) PA-enhanced methods are robust across different corruption types and severity levels; (iii) PA remains effective for nonlinear classifiers with different activation functions; and (iv) combining PA with aggregation can achieve near-perfect agreement with predictions obtained without randomization under suitable privacy budgets and aggregation sizes.

7 Limitations

Our theoretical analysis establishes the optimal basis and scale allocation for the noise-induced component of the first-order utility surrogate under the considered geometric constraint, and further characterizes the relation between the first-order surrogate and nonlinear output distortion and prediction stability. For nonlinear downstream tasks, the end-to-end privacy guarantee remains exact, whereas the utility improvement over the isotropic baseline is established empirically rather than through a global guarantee on the nonlinear task loss. A remaining gap is that the transformation-dependent clipping term and the resulting nonlinear task loss are not jointly optimized with the noise allocation, as such joint optimization requires strong structural assumptions. Since these effects depend on both the downstream model geometry and the representation distribution, we leave the development of tractable end-to-end objectives to future work on specific task settings under such assumptions.

AI USE STATEMENT

In this work, we used generative AI tools to refine the wording and improve the readability of the manuscript, identify potentially missing references or gaps in the related work discussion, and check experimental implementations for potential errors. The authors reviewed all AI-generated suggestions before incorporating them into the manuscript or experimental code. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The authors believe that this paper submission raises no questions regarding the Code of Ethics.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we provide the theoretical analysis and proofs in Appendix A.4 and the privacy guarantees and their proofs in Appendix A.6. The experimental environment and shared hyperparameters are described in Appendix A.10.1. Dataset construction, public/private data splits, and evaluation protocols are detailed in Appendices A.10.2 and A.10.3, while downstream model architectures and feature extraction procedures are described in Appendices A.10.4 and A.10.5. Additional experimental results and sensitivity analyses are provided in Appendices A.10.6–A.10.11. Upon acceptance, we will publicly release our source code together with a README containing instructions for environment setup and reproduction of the experimental results.

REFERENCES

Mohammad Alaggan, Sébastien Gambs, and Anne-Marie Kermarrec. Heterogeneous Differential Privacy. *Journal of Privacy and Confidentiality*, 7(2):127–158, 2016.

Apple Differential Privacy Team. Learning with Privacy at Scale. Technical report, Apple Machine Learning Journal, 2017. URL `https://machinelearning.apple.com/research/learning-with-privacy-at-scale`.

Hilal Asi, Vitaly Feldman, Tomer Koren, and Kunal Talwar. Optimal Algorithms for Mean Estimation under Local Differential Privacy. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 1046–1056, 2022.

Martin Aumüller, Christian Janos Lebeda, Boel Nelson, and Rasmus Pagh. PLAN: Variance-Aware Private Mean Estimation. *Proceedings on Privacy Enhancing Technologies*, 2024(3):606–625, 2024.

Borja Balle and Yu-Xiang Wang. Improving the Gaussian Mechanism for Differential Privacy: Analytical Calibration and Optimal Denoising. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pages 394–403, 2018.

Abhishek Bhowmick, John C. Duchi, Julien Freudiger, Gaurav Kapoor, and Ryan Rogers. Protection Against Reconstruction and Its Applications in Private Federated Learning. *arXiv preprint arXiv:1812.00984*, 2018.

Jiangnan Cheng, Ao Tang, and Sandeep Chinchali. Task-aware Privacy Preservation for Multi-dimensional Data. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, volume 162, pages 3835–3851, 2022.

Paul G. Constantine, Eric Dow, and Qiqi Wang. Active Subspace Methods in Theory and Practice: Applications to Kriging Surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

Graham Cormode, Tejas Kulkarni, and Divesh Srivastava. Answering Range Queries Under Local Differential Privacy. *Proceedings of the VLDB Endowment*, 12(10):1126–1138, 2019.

Jean-Michel D. Smart Meters in London. `https://www.kaggle.com/datasets/jeanmidev/smart-meters-in-london`, 2019. Accessed: 2026.

Yuval Dagan, Michael I. Jordan, Xuelin Yang, Lydia Zakynthinou, and Nikita Zhivotovskiy. Dimension-free Private Mean Estimation for Anisotropic Distributions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 120667–120698, 2024.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. Collecting telemetry data privately. In *Advances in Neural Information Processing Systems (NIPS)*, volume 30, pages 3574–3583, 2017.

Jiawei Duan, Qingqing Ye, and Haibo Hu. Utility Analysis and Enhancement of LDP Mechanisms in High-Dimensional Space. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, pages 407–419, 2022.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 429–438, 2013.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Minimax Optimal Procedures for Locally Private Estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the Third Conference on Theory of Cryptography (TCC)*, Lecture Notes in Computer Science, pages 265–284. Springer-Verlag, 2006.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 1054–1067, 2014.

Huiyu Fang, Liquan Chen, Yali Liu, and Yuan Gao. Locally Differentially Private Frequency Estimation Based on Convolution Framework. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 2208–2222, 2023.

Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the Accuracy of Differentially Private Histograms Through Consistency. *Proceedings of the VLDB Endowment*, 3(1): 1021–1032, 2010.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning For Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Charlie Hou, Akshat Shrivastava, Hongyuan Zhan, Rylan Conway, Trang Le, Adithya Sagar, Giulia Fanti, and Daniel Lazar. PrE-Text: Training Language Models on Private Federated Data in the Age of LLMs. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 19043–19061, 2024.

Ziyue Huang, Yuting Liang, and Ke Yi. Instance-Optimal Mean Estimation Under Differential Privacy. In *Advances in Neural Information Processing Systems (NIPS)*, volume 34, pages 25993–26004, 2021.

Jinyuan Jia and Neil Zhenqiang Gong. Calibrate: Frequency Estimation and Heavy Hitter Identification with Local Differential Privacy via Incorporating Prior Knowledge. In *IEEE INFOCOM 2019 - IEEE Conference on Computer Communications*, pages 2008–2016, 2019.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete Distribution Estimation under Local Privacy. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, volume 48, pages 2436–2444, 2016.

Shiva Prasad Kasiviswanathan, Homin K Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What Can We Learn Privately? *SIAM Journal on Computing*, 40(3):793–826, 2011.

Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A Framework for Mathematical Privacy Definitions. *ACM Transactions on Database Systems*, 39(1):3:1–3:36, 2014.

Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto, 2009.

Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward Training at ImageNet Scale with Differential Privacy. *arXiv preprint arXiv:2201.12328*, 2022.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-Based Learning Applied to Document Recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Peihua Mai, Ran Yan, Zhe Huang, Youjia Yang, and Yan Pang. Split-and-Denoise: Protect Large Language Model Inference with Local Differential Privacy. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 34281 – 34302, 2024.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Prakash Ramrakhyani, Ali Jalali, Dean Tullsen, and Hadi Esmaeilzadeh. Shredder: Learning Noise Distributions to Protect Inference Privacy. In *Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems*, page 3–18, 2020.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. Not All Features Are Equal: Discovering Essential Features for Preserving Prediction Privacy. In *Proceedings of the Web Conference 2021*, page 669–680, New York, NY, USA, 2021.

Gokularam Muthukrishnan and Sheetal Kalyani. Differential Privacy With Higher Utility by Exploiting Coordinate-Wise Disparity: Laplace Mechanism Can Beat Gaussian in High Dimensions. *IEEE Transactions on Information Forensics and Security*, 20:2836–2851, 2025.

Milad Nasr, Saeed Mahloujifar, Xinyu Tang, Prateek Mittal, and Amir Houmansadr. Effectively Using Public Data in Privacy Preserving Machine Learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 25718–25732, 2023.

NhatHai Phan, Xintao Wu, Han Hu, and Dejing Dou. Adaptive Laplace Mechanism: Differential Privacy Preservation in Deep Learning. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 385–394, 2017.

Caihong Qin and Yang Bai. Classification under local differential privacy with model reversal and model averaging. *Journal of Machine Learning Research*, 27(5):1–44, 2026. URL `http://jmlr.org/papers/v27/24-0290.html`.

Sina Sajadmanesh and Daniel Gatica-Perez. Locally Private Graph Neural Networks. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 2130–2145, 2021.

UK Power Networks. Smartmeter energy consumption data in london households. URL `https://data.london.gov.uk/dataset/smartmeter-energy-use-data-in-london-households`. Low Carbon London project data.

Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Cheung Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and Analyzing Multidimensional Data with Local Differential Privacy. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 638–649, 2019.

Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally Differentially Private Protocols for Frequency Estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, 2017.

Tianhao Wang, Milan Lopuhaä-Zwakenberg, Zitao Li, Boris Skoric, and Ninghui Li. Locally Differentially Private Frequency Estimation with Consistency. In *Proceedings of the 27th Network and Distributed System Security Symposium (NDSS)*, 2020.

A Appendix

A.1 Related Work

Task-Aware Noise for Private Remote Inference. A line of work on privacy-preserving remote inference perturbs or suppresses the data sent to a cloud service while seeking to preserve predictive accuracy. Shredder (Mireshghallah et al., 2020) splits inference between the edge device and the cloud, learning additive noise distributions for intermediate representations through offline optimization. Cloak (Mireshghallah et al., 2021) identifies task-essential features via gradient-based perturbation maximization and suppresses the rest. While these approaches share our intuition that task-irrelevant components can tolerate substantial perturbation and similarly release representations rather than model outputs, they rely on data-driven optimization and lack formal differential privacy guarantees. Meanwhile, Split-and-Denoise (Mai et al., 2024) perturbs client-side token embeddings under metric local differential privacy (d_χ-privacy) before server-side LLM inference, and the client denoises the returned outputs to obtain the final result. However, it fundamentally differs from our scope in that it is designed for d_χ-privacy and it requires the private data owner to perform denoising locally. In contrast, our approach provides strict (ϵ, δ)-LDP guarantees for representation release, deriving the noise geometry in closed form from the Jacobian of the downstream model and operating in a rotated task-aligned basis. Moreover, since no client involvement is required after model inference on the server, our approach incurs no additional communication rounds.

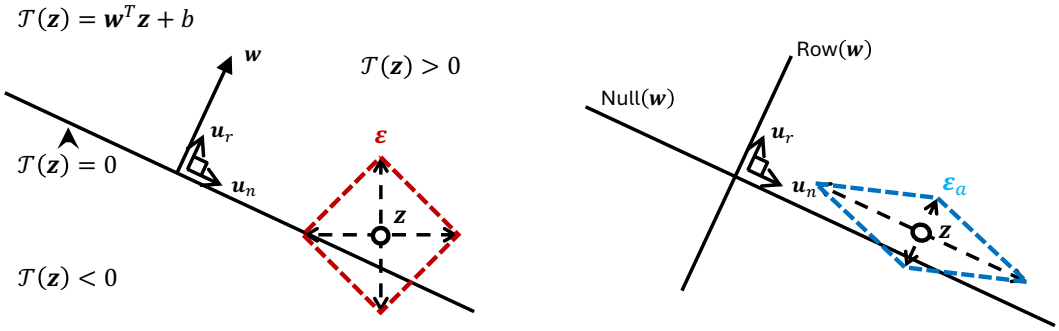
Relevance-Aware. A line of related work dynamically allocates privacy budgets based on feature importance. For instance, the Adaptive Laplace Mechanism (AdLM) (Phan et al., 2017) employs Layer-wise Relevance Propagation (LRP) to assign larger privacy budgets to features that contribute more to the final prediction, thereby reducing noise on highly relevant dimensions. While both AdLM and our method share a task-aware intuition, they differ fundamentally in their mathematical foundations. AdLM relies on LRP to compute scalar attribution scores for individual features. In contrast, our approach leverages the Jacobian matrix to evaluate the sensitivity of the output with respect to the input. This mathematically grounded approach allows us to analytically decompose the input space into geometrically meaningful structures, i.e., the local row and null spaces, rather than merely assigning scalar weights to individual pixels.

Correlation-Aware. Several studies address DP vulnerabilities when the assumption of record independence is violated. For instance, Pufferfish (Kifer and Machanavajjhala, 2014) incorporates joint probability distributions directly into the privacy definition to account for data dependencies. Similarly, PLAN (Aumüller et al., 2024) injects anisotropic noise by allocating the privacy budget based on the data geometry inherent in a known covariance matrix, a method later extended to unknown covariances by (Dagan et al., 2024). While these works (Kifer and Machanavajjhala, 2014; Aumüller et al., 2024; Dagan et al., 2024) explore anisotropic noise injection guided solely by intrinsic data correlations, our approach fundamentally differs by scaling noise based on the task-dependent importance of subspaces utilized by the downstream model.

Feature-Wise DP. Alaggan et al. (2016) introduced the Heterogeneous Differential Privacy (HDP), which captures the non-uniformity of privacy expectations both among different users and across various data items belonging to the same individual. HDP enables item-grained privacy, allowing for the adaptive calibration of noise levels based on the specific sensitivity of each piece of information. Our approach primarily differs from HDP in its treatment of information sensitivity. Unlike HDP, which accounts for non-uniform sensitivity, we use an isotropically calibrated base randomizer and reshape its effective perturbation across task directions. Furthermore, we enhance overall utility by applying a bijective transformation that reduces the noise injected into key dimensions.

Post-Processing. Enhancing utility under LDP often relies on post-processing techniques such as data aggregation, denoising, and statistical inference. Foundational mechanisms, including industry deployments by Google (Erlingsson et al., 2014), Apple (Apple Differential Privacy Team, 2017), and Microsoft (Ding et al., 2017), as well as variance-optimized encodings like OUE/OLH (Wang et al., 2017), established standard aggregation pipelines. To further reduce estimation errors, recent works have incorporated structural priors, such as hierarchical consistency (Hay et al., 2010; Cormode et al., 2019), probability simplex constraints (Wang et al., 2020), iterative calibration (Jia and Gong, 2019), and signal-processing filters (Fang et al., 2023).

A.2 Motivating Example



(a) Linear Binary Classification in $\mathbb{R}^2$ (b) Anisotropically Reshaping Noise in $\mathbb{R}^2$

Figure 1: **Motivating Example.** $\mathcal{T}(z)$ is a linear classifier with normal vector w orthogonal to the decision boundary ($\mathcal{T}(z) = 0$). u_r and u_n are basis vectors for the row and null spaces, respectively. Dotted arrows denote perturbations around z. The red and blue dashed lines describe the shape of the Laplace noise. Notably, perturbations along u_r may alter the prediction, while those along u_n do not. **(a) Isotropic noise.** Isotropic noise ξ is generated. **(b) Anisotropic Reshaping.** Isotropic noise ξ is reshaped into ξ_a by attenuating the u_r axis and amplifying the u_n axis.

A.3 Proposed Approach: Description

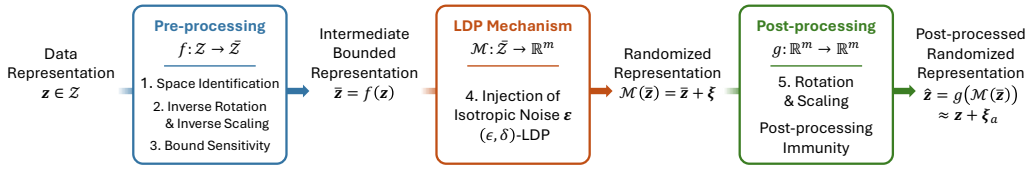


Figure 2: Overview of the proposed approach.

We provide a more detailed explanation of the procedure. Here, we primarily focus on the row space for clarity. Let $\lambda_r \in (0, 1]$ be a parameter that governs the relationship between the desired variance σ_r^2 of the noise projected into the row space via anisotropic noise ξ_a, and the variance σ^2 of the isotropic noise ξ, such that $\lambda_r = (\frac{\sigma_r}{\sigma})^2$. The procedure consists of five steps:

1. Our approach identifies the row and null spaces based on Jacobian matrices,
2. It inversely rotates and expands the representation in the row space by a factor of $\frac{1}{\sqrt{\lambda_r}}$,
3. It bounds the row space sensitivity,
4. It injects isotropic noise guaranteeing ϵ-LDP, and
5. It rotates and scales the randomized representation down by a factor of $\sqrt{\lambda_r}$, thereby reshaping the isotropic noise into anisotropic noise and restoring the original scale and orientation of the representation.

The pre-processing function f handles the first three steps, and the post-processing function g manages the final step. In the pre-processing step, while space identification and representation alignment are important, bounding sensitivity also plays a crucial role in reducing the noise injected into the row space. Although bounding incurs utility loss, we assume that it is outweighed by the utility degradation caused by ϵ-LDP randomization under high sensitivity (see Section 6.2 for details). In the fourth step, we inject noise via the conventional mechanism to guarantee ϵ-LDP for the predetermined sensitivity, which is proportional to the bounding threshold. In the final step, g applies a fixed invertible affine transformation. The privacy guarantee is preserved due to the post-processing immunity of differential privacy (Dwork and Roth, 2014).

Note that in our approach, noise control is achieved through two key parameters: the bounding threshold $\rho \in \mathbb{R}^+$ and the ratio λ_r of the coordinate-wise noise to the isotropic noise. The magnitude of isotropic noise is determined by the privacy budget, the threshold, and the chosen randomization mechanism. Finally, through the linear bijective rotation and scaling applied during post-processing, noise satisfying ϵ-LDP is automatically distributed across all coordinates.

A.4 Theoretical Analysis

We provide our analysis under the assumption of ℓ_1 radial clipping and its corresponding geometric gauge $\sum_{i=1}^m \lambda_i^{-1/2} = m$. The extension to general ℓ_p clipping and the gauge $\sum_{i=1}^m \lambda_i^{-p/2} = m$ can be derived analogously.

Definition 1 (End-to-end perturbation for additive mechanisms)**.** *Radial ℓ_1 clipping in the transformed space scales $\boldsymbol{L}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})$ by*

$$\theta_1(\boldsymbol{z}) = \begin{cases} \min\left\{1,\ \dfrac{\rho}{\|\boldsymbol{L}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\|_1}\right\}, & \boldsymbol{L}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) \neq \mathbf{0}, \\ 1, & \boldsymbol{L}^{-1}(\boldsymbol{z}-\boldsymbol{\mu}) = \mathbf{0}. \end{cases}$$

This operation shrinks the centered representation toward $\boldsymbol{\mu}$ while preserving the direction of $\boldsymbol{z}-\boldsymbol{\mu}$.

For an additive mechanism, the i-th randomized representation can be written as

$$\begin{aligned} \hat{\boldsymbol{z}}_i &= \boldsymbol{L}\left[\theta_1(\boldsymbol{z}_i)\left\{\boldsymbol{L}^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu})\right\} + \boldsymbol{\xi}\right] + \boldsymbol{\mu} \\ &= \theta_1(\boldsymbol{z}_i)\boldsymbol{L}\left\{\boldsymbol{L}^{-1}(\boldsymbol{z}_i-\boldsymbol{\mu})\right\} + \boldsymbol{L}\boldsymbol{\xi} + \boldsymbol{\mu} \\ &= \theta_1(\boldsymbol{z}_i)\left(\boldsymbol{z}_i-\boldsymbol{\mu}\right) + \boldsymbol{L}\boldsymbol{\xi} + \boldsymbol{\mu} \end{aligned}$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$.

Note that the scaling factor of the ℓ_1-ball clipping operation evaluates to a scalar given a representation $\boldsymbol{z}$ and the parameters $\{\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{\mu}, p, \rho\}$.

Therefore, the end-to-end perturbation is

$$\boldsymbol{h}_i = \hat{\boldsymbol{z}}_i - \boldsymbol{z}_i = \boldsymbol{d}_i + \boldsymbol{L}\boldsymbol{\xi},$$

where

$$\boldsymbol{d}_i = \big(\theta_1(\boldsymbol{z}_i) - 1\big)(\boldsymbol{z}_i - \boldsymbol{\mu}).$$

Here, $\boldsymbol{d}_i = \mathbf{0}$ for an unclipped sample and points toward $\boldsymbol{\mu}$ otherwise.

Definition 2 (Empirical Jacobian Gram second-moment matrix)**.** *Define*

$$\widehat{\boldsymbol{\Omega}} = \boldsymbol{J}^\top \boldsymbol{J} = \frac{1}{N}\sum_{i=1}^N \boldsymbol{J}_i^\top \boldsymbol{J}_i.$$

Let its eigendecomposition be

$$\widehat{\boldsymbol{\Omega}} = \boldsymbol{V}\,\mathrm{diag}(\gamma_1, \ldots, \gamma_m)\boldsymbol{V}^\top,$$

where $\gamma_1 \geq \cdots \geq \gamma_m \geq 0$.

Definition 3 (Empirical first-order risk surrogate)**.** *Define*

$$\widehat{R}_{\mathrm{FO}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{J}_i\boldsymbol{h}_i\|_2^2\right].$$

Using $\boldsymbol{h}_i = \boldsymbol{d}_i + \boldsymbol{L}\boldsymbol{\xi}$, we obtain

$$\widehat{R}_{\mathrm{FO}} = \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{J}_i\boldsymbol{d}_i + \boldsymbol{J}_i\boldsymbol{L}\boldsymbol{\xi}\|_2^2\right].$$

Because $\boldsymbol{d}_i$ is fixed conditional on $\boldsymbol{z}_i$ and $\mathbb{E}[\boldsymbol{\xi}] = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\xi}) = \sigma^2\boldsymbol{I}$, the cross term vanishes, yielding

$$\widehat{R}_{\mathrm{FO}} = \underbrace{\sigma^2\,\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{U}^\top\widehat{\boldsymbol{\Omega}}\boldsymbol{U}\right)}_{\widehat{R}_{\mathrm{FO,Noise}}} + \underbrace{\frac{1}{N}\sum_{i=1}^N \|\boldsymbol{J}_i\boldsymbol{d}_i\|_2^2}_{\widehat{R}_{\mathrm{FO,Clip}}}.$$

The following propositions optimize $\widehat{R}_{\mathrm{FO,Noise}}$, not the clipping term.

Definition 4 (Empirical nonlinear risk)**.** *For a nonlinear task model $\mathcal{T}$, we define the empirical nonlinear (ENL) risk as*

$$\widehat{R}_{\mathrm{NL}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}}\left[\|\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i)\|_2^2\right].$$

Proposition 1 (Optimality of the Jacobian basis for the noise component)**.** *Let*

$$\widehat{\boldsymbol{\Omega}} = \boldsymbol{V}\,\mathrm{diag}(\gamma_1, \ldots, \gamma_m)\boldsymbol{V}^\top, \qquad \gamma_1 \geq \cdots \geq \gamma_m \geq 0.$$

Fix an ordered positive scale spectrum $0 < \lambda_1 \leq \cdots \leq \lambda_m$ and define $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$. Then

$$\min_{\boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}} \mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{U}^\top\widehat{\boldsymbol{\Omega}}\boldsymbol{U}\right) = \sum_{i=1}^{m}\lambda_i\gamma_i.$$

One minimizer is $\boldsymbol{U} = \boldsymbol{V}$, with the largest Jacobian-Gram eigenvalues paired with the smallest scales. Equivalently, for a fixed ordered spectrum, the Jacobian right-singular-vector basis is optimal over all orthogonal bases for the noise-induced first-order objective. Non-uniqueness may arise from sign changes and transformations within repeated-eigenvalue or equal-scale blocks.

Proof. Let $\boldsymbol{H} = \boldsymbol{V}^\top\boldsymbol{U}$, which is orthogonal. Then

$$\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{U}^\top\widehat{\boldsymbol{\Omega}}\boldsymbol{U}\right) = \mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{H}^\top\,\mathrm{diag}(\gamma_1, \ldots, \gamma_m)\boldsymbol{H}\right) = \sum_{i,j}\lambda_i\gamma_j H_{ji}^2.$$

Define $G_{ij} = H_{ji}^2$. Because $\boldsymbol{H}$ is orthogonal, $\boldsymbol{G}$ is doubly stochastic. Relaxing the feasible set of such orthostochastic matrices to the full Birkhoff polytope therefore provides a lower bound.

The relaxed objective is linear in $\boldsymbol{G}$, so its minimum is attained at an extreme point of the Birkhoff polytope, namely a permutation matrix. By the rearrangement inequality, the minimizing permutation pairs the largest γ_i with the smallest λ_i, giving $\sum_{i=1}^{m}\lambda_i\gamma_i$.

This lower bound is attainable in the original feasible set because permutation matrices are orthogonal. Under the stated ordering, $\boldsymbol{H} = \boldsymbol{I}$, equivalently $\boldsymbol{U} = \boldsymbol{V}$, attains the bound. Hence the relaxation is tight. □

Remark 1 (Rank deficiency and the finite cap)**.** *When $\gamma_i = 0$, the noise objective does not depend directly on λ_i. If the scales are optimized under $\sum_{j=1}^{m}\lambda_j^{-1/2} = m$ without an upper bound, the infimum sends the corresponding scale toward $\lambda_i \to \infty$. Equivalently, $t_i = \lambda_i^{-1/2} \to 0$, which leaves more normalization budget for directions with positive γ_i. A finite optimizer therefore need not exist in an exact null direction.*

The cap $\lambda_i \leq \lambda_{\max}$ removes this degeneracy. For a near-zero but positive γ_i, the uncapped optimal scale is finite but may be very large and sensitive to estimation error. In that case, the cap additionally limits extreme anisotropy and improves numerical robustness.

Proposition 2 (Joint optimality of the Jacobian basis and scale allocation)**.** *Assume $1 < \lambda_{\max} < \infty$ and $\widehat{\boldsymbol{\Omega}} \neq \boldsymbol{0}$. Consider*

$$\min_{\substack{\boldsymbol{U}^\top\boldsymbol{U}=\boldsymbol{I}\\ 0<\lambda_i\leq\lambda_{\max}}} \sigma^2\,\mathrm{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{U}^\top\widehat{\boldsymbol{\Omega}}\boldsymbol{U}\right) \qquad \textit{subject to} \qquad \sum_{i=1}^{m}\lambda_i^{-1/2} = m.$$

A simultaneous permutation of the columns of $\boldsymbol{U}$ and the diagonal entries of $\boldsymbol{\Lambda}$ does not change the optimization problem. We may therefore assume, without loss of generality, that $0 < \lambda_1 \leq \cdots \leq \lambda_m$. By Proposition 1, the minimization over $\boldsymbol{U}$ is attained by the Jacobian basis and reduces the joint problem to

$$\min_{\boldsymbol{\Lambda}} \quad \sigma^2\sum_{i=1}^{m}\gamma_i\lambda_i.$$

Set $t_i = \lambda_i^{-1/2}$ and $t_{\min} = \lambda_{\max}^{-1/2}$. The problem becomes

$$\min_{\boldsymbol{t}} \quad \sum_{i=1}^{m}\gamma_i t_i^{-2} \qquad \textit{subject to} \qquad \sum_{i=1}^{m} t_i = m, \qquad t_i \geq t_{\min}.$$

This is a convex optimization problem. Its solution has the water-filling form

$$t_i^\star = \max\left\{t_{\min}, \alpha\gamma_i^{1/3}\right\}, \qquad \lambda_i^\star = (t_i^\star)^{-2},$$

where $\alpha > 0$ is chosen so that

$$\sum_{i=1}^m t_i^\star = \sum_{i=1}^m \max\left\{t_{\min}, \alpha\gamma_i^{1/3}\right\} = m.$$

Let the resulting capped and free sets be

$$C = \left\{i : \alpha\gamma_i^{1/3} \le t_{\min}\right\}, \qquad F = C^c.$$

The normalization constraint then gives $|C|t_{\min} + \alpha\sum_{j\in F}\gamma_j^{1/3} = m$. Therefore,

$$\alpha = \frac{m - |C|t_{\min}}{\sum_{j\in F}\gamma_j^{1/3}} = \frac{m - |C|\lambda_{\max}^{-1/2}}{\sum_{j\in F}\gamma_j^{1/3}}.$$

Consequently, the optimal scales can be written as

$$\lambda_i^\star = \begin{cases} \lambda_{\max}, & i \in C, \\ \left(\dfrac{m - |C|\lambda_{\max}^{-1/2}}{\sum_{j\in F}\gamma_j^{1/3}}\gamma_i^{1/3}\right)^{-2}, & i \in F. \end{cases}$$

Thus, on the free coordinates, $\lambda_i^\star \propto \gamma_i^{-2/3}$, whereas zero and sufficiently small singular-value directions receive the finite cap. The resulting scale spectrum is ordered oppositely to the Jacobian-Gram eigenvalues, as required by Proposition 1.

When no coordinate is capped, $C = \varnothing$ and $F = \{1, \dots, m\}$, so $\alpha = \frac{m}{\sum_{j=1}^m \gamma_j^{1/3}}$ and

$$\lambda_i^\star = \left(\frac{m\,\gamma_i^{1/3}}{\sum_{j=1}^m \gamma_j^{1/3}}\right)^{-2}.$$

Proof. The Lagrangian is

$$\mathcal{L} = \sum_i \gamma_i t_i^{-2} + \eta\left(\sum_i t_i - m\right) + \sum_i \nu_i(t_{\min} - t_i),$$

where $\nu_i \ge 0$. The stationarity condition is

$$-2\gamma_i t_i^{-3} + \eta - \nu_i = 0.$$

For a free coordinate, $\nu_i = 0$, so

$$t_i = \left(\frac{2\gamma_i}{\eta}\right)^{1/3} = \alpha\gamma_i^{1/3}.$$

For a capped coordinate, $t_i = t_{\min}$, and complementary slackness gives

$$\nu_i = \eta - 2\gamma_i t_{\min}^{-3} \ge 0.$$

Hence capping occurs for sufficiently small γ_i. Enforcing the normalization constraint gives

$$\alpha = \frac{m - |C|t_{\min}}{\sum_{j\in F}\gamma_j^{1/3}}.$$

Because $\lambda_{\max} > 1$ and at least one $\gamma_i > 0$, the equation determining α has a unique solution. Strict convexity gives uniqueness on coordinates with $\gamma_i > 0$. For $\gamma_i = 0$, setting $t_i = t_{\min}$, equivalently $\lambda_i = \lambda_{\max}$, leaves the largest possible normalization budget for the positive-singular-value coordinates and is therefore optimal. □

Proposition 3 (Remainder bound for the empirical first-order risk surrogate)**.** *Let $\mathcal{T}$ be a piecewise-affine ReLU network that is globally K-Lipschitz and differentiable at each $\boldsymbol{z}_i$. For each i, let $\mathcal{R}_i$ be a closed convex polyhedral activation cell containing $\boldsymbol{z}_i$, on which $\mathcal{T}$ is affine with Jacobian $\boldsymbol{J}_i$, and define the nonlinear remainder $\boldsymbol{\zeta}_i$ by*

$$\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i) = \boldsymbol{J}_i \boldsymbol{h}_i + \boldsymbol{\zeta}_i.$$

Let $\mathcal{A}_i$ denote the event

$$\mathcal{A}_i = \{\boldsymbol{z}_i + t\boldsymbol{h}_i \in \mathcal{R}_i \ \textit{for all } t \in [0,1]\},$$

and define

$$\bar{P} = \frac{1}{N}\sum_{i=1}^{N} \Pr_{\boldsymbol{\xi}}(\mathcal{A}_i^c).$$

Since $\mathcal{R}_i$ is convex and $\boldsymbol{z}_i \in \mathcal{R}_i$, the event $\mathcal{A}_i$ is equivalently characterized by its endpoint, $\mathcal{A}_i = \{\boldsymbol{z}_i + \boldsymbol{h}_i \in \mathcal{R}_i\}$. On $\mathcal{A}_i$ the map $\mathcal{T}$ agrees with its affine restriction on $\mathcal{R}_i$, so that $\boldsymbol{\zeta}_i = \mathbf{0}$; the first-order expression is exact whenever the perturbation stays within $\mathcal{R}_i$.

Assume that

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{h}_i\|_2^4\right] < \infty \qquad \textit{for all } i \in \{1,\dots,N\}.$$

Then

$$\left|\sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}}\right| \le 2K\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{h}_i\|_2^2 \mathbf{1}_{\mathcal{A}_i^c}\right]\right)^{1/2},$$

and consequently,

$$\left|\sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}}\right| \le 2K\left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{h}_i\|_2^4\right]\right)^{1/4}\bar{P}^{1/4}.$$

Proof. For a sequence of random vectors $\{\boldsymbol{v}_i\}_{i=1}^N$, define

$$\left\|\{\boldsymbol{v}_i\}_{i=1}^N\right\|_{\mathcal{H}} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{v}_i\|_2^2\right]\right)^{1/2}.$$

This is the root-mean-square norm with respect to the joint average over the empirical index i (uniform on $\{1,\dots,N\}$) and the randomness of $\boldsymbol{\xi}$. Under this norm,

$$\sqrt{\widehat{R}_{\mathrm{NL}}} = \left\|\{\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i)\}_{i=1}^N\right\|_{\mathcal{H}}, \qquad \sqrt{\widehat{R}_{\mathrm{FO}}} = \left\|\{\boldsymbol{J}_i\boldsymbol{h}_i\}_{i=1}^N\right\|_{\mathcal{H}}.$$

Since $\mathcal{T}$ is globally K-Lipschitz and differentiable at $\boldsymbol{z}_i$, its Jacobian satisfies

$$\|\boldsymbol{J}_i\|_{\mathrm{op}} = \sup_{\boldsymbol{u}\neq\mathbf{0}}\frac{\|\boldsymbol{J}_i\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} \le K,$$

so both $\|\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i)\|_2$ and $\|\boldsymbol{J}_i\boldsymbol{h}_i\|_2$ are bounded by $K\|\boldsymbol{h}_i\|_2$. Since $\mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{h}_i\|_2^4] < \infty$ implies $\mathbb{E}_{\boldsymbol{\xi}}[\|\boldsymbol{h}_i\|_2^2] < \infty$, both $\widehat{R}_{\mathrm{NL}}$ and $\widehat{R}_{\mathrm{FO}}$ are finite.

By the reverse triangle inequality,

$$\left|\sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}}\right| \le \left\|\{\boldsymbol{\zeta}_i\}_{i=1}^N\right\|_{\mathcal{H}} = \left(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\boldsymbol{\xi}}\left[\|\boldsymbol{\zeta}_i\|_2^2\right]\right)^{1/2}.$$

For the remainder itself, the triangle inequality together with the two bounds above gives the deterministic worst case

$$\|\boldsymbol{\zeta}_i\|_2 = \|\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i) - \boldsymbol{J}_i\boldsymbol{h}_i\|_2 \le \|\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i)\|_2 + \|\boldsymbol{J}_i\boldsymbol{h}_i\|_2 \le 2K\|\boldsymbol{h}_i\|_2.$$

Since $\boldsymbol{\zeta}_i = \mathbf{0}$ on $\mathcal{A}_i$, this worst case is only active on the complement:

$$\|\boldsymbol{\zeta}_i\|_2^2 \le 4K^2 \|\boldsymbol{h}_i\|_2^2 \mathbf{1}_{\mathcal{A}_i^c}.$$

Substituting into the preceding reverse-triangle bound gives

$$\left| \sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}} \right| \le 2K \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^2 \mathbf{1}_{\mathcal{A}_i^c} \right] \right)^{1/2}.$$

Finally, since $\|\cdot\|_{\mathcal{H}}$ is the root-mean-square norm for this joint average, the Cauchy–Schwarz inequality applies to the pair $\|\boldsymbol{h}_i\|_2^2$ and $\mathbf{1}_{\mathcal{A}_i^c}$ over the same average; using $\mathbf{1}_{\mathcal{A}_i^c}^2 = \mathbf{1}_{\mathcal{A}_i^c}$,

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^2 \mathbf{1}_{\mathcal{A}_i^c} \right] &\le \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^4 \right] \right)^{1/2} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\mathbf{1}_{\mathcal{A}_i^c} \right] \right)^{1/2} \\ &= \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^4 \right] \right)^{1/2} \bar{P}^{1/2}. \end{aligned}$$

Taking the outer square root yields

$$\left| \sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}} \right| \le 2K \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^4 \right] \right)^{1/4} \bar{P}^{1/4},$$

which completes the proof. □

Corollary 1 (Direct bound on the empirical nonlinear-risk discrepancy)**.** *Under the assumptions of Proposition 3, define*

$$B = 2K \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{\xi}} \left[\|\boldsymbol{h}_i\|_2^4 \right] \right)^{1/4} \bar{P}^{1/4}.$$

Then

$$\left| \widehat{R}_{\mathrm{NL}} - \widehat{R}_{\mathrm{FO}} \right| \le 2B \sqrt{\widehat{R}_{\mathrm{FO}}} + B^2.$$

Proof. By Proposition 3,

$$\left| \sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}} \right| \le B.$$

Hence,

$$\left| \widehat{R}_{\mathrm{NL}} - \widehat{R}_{\mathrm{FO}} \right| = \left| \sqrt{\widehat{R}_{\mathrm{NL}}} - \sqrt{\widehat{R}_{\mathrm{FO}}} \right| \left(\sqrt{\widehat{R}_{\mathrm{NL}}} + \sqrt{\widehat{R}_{\mathrm{FO}}} \right) \le B \left(2\sqrt{\widehat{R}_{\mathrm{FO}}} + B \right),$$

where the last inequality follows from

$$\sqrt{\widehat{R}_{\mathrm{NL}}} \le \sqrt{\widehat{R}_{\mathrm{FO}}} + B.$$

This proves the result. □

Proposition 4 (Deterministic bound for the empirical noise-risk discrepancy)**.** *Let $\boldsymbol{\Omega}$ denote the population Jacobian second-moment matrix and let $\widehat{\boldsymbol{\Omega}}$ denote its empirical estimate obtained from N public samples.*

Define

$$R_{\mathrm{FO,Noise}} = \sigma^2 \operatorname{Tr}\left(\boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{\Omega} \boldsymbol{U} \right)$$

and

$$\widehat{R}_{\mathrm{FO,Noise}} = \sigma^2 \operatorname{Tr}\left(\boldsymbol{\Lambda} \boldsymbol{U}^\top \widehat{\boldsymbol{\Omega}} \boldsymbol{U} \right).$$

Then,

$$\left| R_{\mathrm{FO,Noise}} - \widehat{R}_{\mathrm{FO,Noise}} \right| = \sigma^2 \left| \operatorname{Tr}\left(\boldsymbol{\Lambda} \boldsymbol{U}^\top \left(\boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}} \right) \boldsymbol{U} \right) \right| \le \sigma^2 \left\| \boldsymbol{\Omega} - \widehat{\boldsymbol{\Omega}} \right\|_{\mathrm{op}} \operatorname{Tr}(\boldsymbol{\Lambda}).$$

Proof. Using cyclic invariance of the trace,

$$\operatorname{Tr}\left(\boldsymbol{\Lambda}\boldsymbol{U}^\top\left(\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}\right)\boldsymbol{U}\right)=\operatorname{Tr}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\left(\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}\right)\right).$$

By the duality between the nuclear and operator norms,

$$\left|\operatorname{Tr}\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\left(\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}\right)\right)\right|\le\left\|\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right\|_*\|\left(\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}\right)\|_{\mathrm{op}}.$$

Since $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Lambda}\succeq 0$,

$$\left\|\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top\right\|_*=\operatorname{Tr}(\boldsymbol{\Lambda}).$$

Multiplying by σ^2 gives the result. □

Corollary 2 (Finite-sample convergence of the empirical noise risk)**.** *Suppose that the public samples are independently and identically distributed and that the Jacobian second-moment estimator satisfies*

$$\left\|\boldsymbol{\Omega}-\widehat{\boldsymbol{\Omega}}\right\|_{\mathrm{op}}=\mathcal{O}_P\left(\sqrt{\frac{\log m}{N}}\right),$$

as obtained, for example, under appropriate boundedness or sub-Gaussian assumptions through a matrix concentration inequality.

Then,

$$\left|R_{\mathrm{FO,Noise}}-\widehat{R}_{\mathrm{FO,Noise}}\right|=\mathcal{O}_P\left(\sigma^2\operatorname{Tr}(\boldsymbol{\Lambda})\sqrt{\frac{\log m}{N}}\right).$$

Proposition 5 (Public–private discrepancy bound for the empirical first-order noise risk)**.** *Let* $\boldsymbol{\Omega}_{\mathrm{pub/priv}}$ *denote the Jacobian second-moment matrix under the public/private population.*

Let

$$\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}=\boldsymbol{J}_{\mathrm{pub}}^\top\boldsymbol{J}_{\mathrm{pub}}=\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{J}_{\mathrm{pub},i}^\top\boldsymbol{J}_{\mathrm{pub},i}$$

denote the empirical public Jacobian second-moment matrix.

Define the empirical public first-order noise risk as

$$\widehat{R}_{\mathrm{pub,FO,Noise}}=\sigma^2\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}\right),$$

and define the corresponding private-population noise risk, evaluated using the public basis and scale allocation, as

$$R_{\mathrm{priv,FO,Noise}}=\sigma^2\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\boldsymbol{\Omega}_{\mathrm{priv}}\boldsymbol{U}_{\mathrm{pub}}\right).$$

Then the public–private discrepancy bound for the empirical first-order noise risk is given by

$$\left|R_{\mathrm{priv,FO,Noise}}-\widehat{R}_{\mathrm{pub,FO,Noise}}\right|\le\sigma^2\left\|\boldsymbol{\Omega}_{\mathrm{priv}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right\|_{\mathrm{op}}\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\right).$$

where

$$\left\|\boldsymbol{\Omega}_{\mathrm{priv}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right\|_{\mathrm{op}}\le\left\|\boldsymbol{\Omega}_{\mathrm{priv}}-\boldsymbol{\Omega}_{\mathrm{pub}}\right\|_{\mathrm{op}}+\left\|\boldsymbol{\Omega}_{\mathrm{pub}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right\|_{\mathrm{op}}.$$

Proof. By the definitions above,

$$\left|R_{\mathrm{priv,FO,Noise}}-\widehat{R}_{\mathrm{pub,FO,Noise}}\right|=\sigma^2\left|\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\left(\boldsymbol{\Omega}_{\mathrm{priv}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right)\boldsymbol{U}_{\mathrm{pub}}\right)\right|.$$

Using cyclic invariance of the trace and the duality between the nuclear and operator norms,

$$\left|\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\left(\boldsymbol{\Omega}_{\mathrm{priv}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right)\boldsymbol{U}_{\mathrm{pub}}\right)\right|\le\left\|\boldsymbol{U}_{\mathrm{pub}}\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\right\|_*\left\|\boldsymbol{\Omega}_{\mathrm{priv}}-\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right\|_{\mathrm{op}}.$$

Since $\boldsymbol{U}_{\mathrm{pub}}$ is orthogonal and $\boldsymbol{\Lambda}_{\mathrm{pub}}\succ\mathbf{0}$,

$$\left\|\boldsymbol{U}_{\mathrm{pub}}\boldsymbol{\Lambda}_{\mathrm{pub}}\boldsymbol{U}_{\mathrm{pub}}^\top\right\|_*=\operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\right).$$

This proves the first inequality. The second follows directly from the triangle inequality for the operator norm. □

Corollary 3 (Finite-sample convergence of the empirical public noise risk)**.** *Suppose that the public samples are independently and identically distributed and that, under the stated boundedness or sub-Gaussian Jacobian assumptions,*

$$\left\|\boldsymbol{\Omega}_{\mathrm{pub}} - \widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}\right\|_{\mathrm{op}} = \mathcal{O}_P\left(\sqrt{\frac{\log m}{N}}\right).$$

Then

$$\left|R_{\mathrm{priv,FO,Noise}} - \widehat{R}_{\mathrm{pub,FO,Noise}}\right| \le \sigma^2 \operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\right) \left\|\boldsymbol{\Omega}_{\mathrm{priv}} - \boldsymbol{\Omega}_{\mathrm{pub}}\right\|_{\mathrm{op}} + \mathcal{O}_P\left(\sigma^2 \operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\right)\sqrt{\frac{\log m}{N}}\right).$$

Hence, the finite-sample estimation component decreases at an $N^{-1/2}$-type rate, up to logarithmic dependence on m, whereas the population-level public–private mismatch need not vanish as N increases. In the matched-population case ($\boldsymbol{\Omega}_{\mathrm{priv}} = \boldsymbol{\Omega}_{\mathrm{pub}}$), the bound reduces to

$$\left|R_{\mathrm{priv,FO,Noise}} - \widehat{R}_{\mathrm{pub,FO,Noise}}\right| = \mathcal{O}_P\left(\sigma^2 \operatorname{Tr}\left(\boldsymbol{\Lambda}_{\mathrm{pub}}\right)\sqrt{\frac{\log m}{N}}\right).$$

Proof. The result follows by combining Proposition 5 with the assumed operator-norm concentration rate for $\widehat{\boldsymbol{\Omega}}_{\mathrm{pub}}$. □

Prediction stability for nonlinear $\arg\max$ tasks. We further relate the first-order perturbation analysis to prediction stability for nonlinear classification tasks. For piecewise-affine ReLU networks, the first-order score perturbation is exact whenever the perturbed representation remains in the same activation region. This yields a bound on the prediction-disagreement probability in terms of activation-region crossing and pairwise margin flipping.

For notational convenience, several symbols in this paragraph are reused from other parts of the paper. Such reuse is local to this paragraph, and the symbols below should be interpreted according to the definitions given here.

Consider a differentiable nonlinear score function $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^K$, whose final prediction is obtained by

$$\mathcal{C}(\boldsymbol{z}) = \underset{k\in\{1,\dots,K\}}{\arg\max}\ \mathcal{T}_k(\boldsymbol{z}).$$

For each sample $\boldsymbol{z}_i$, let $c_i = \mathcal{C}(\boldsymbol{z}_i)$ denote its clean prediction. Let $\boldsymbol{J}_i \in \mathbb{R}^{K\times m}$ denote the Jacobian of $\mathcal{T}$ at $\boldsymbol{z}_i$. Using the end-to-end perturbation of Definition 1,

$$\boldsymbol{h}_i = \boldsymbol{d}_i + \boldsymbol{L}\boldsymbol{\xi}, \qquad \boldsymbol{d}_i = \big(\theta_1(\boldsymbol{z}_i) - 1\big)(\boldsymbol{z}_i - \boldsymbol{\mu}), \qquad \boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2},$$

the first-order approximation of the pairwise score gap between the clean predicted class c_i and a competing class k is defined by

$$\begin{aligned}\mathcal{T}_{c_i}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}_k(\boldsymbol{z}_i + \boldsymbol{h}_i) &\approx \big(\mathcal{T}_{c_i}(\boldsymbol{z}_i) + \nabla\mathcal{T}_{c_i}(\boldsymbol{z}_i)^\top \boldsymbol{h}_i\big) - \big(\mathcal{T}_k(\boldsymbol{z}_i) + \nabla\mathcal{T}_k(\boldsymbol{z}_i)^\top \boldsymbol{h}_i\big)\\ &= \big(\mathcal{T}_{c_i}(\boldsymbol{z}_i) - \mathcal{T}_k(\boldsymbol{z}_i)\big) + \big(\nabla\mathcal{T}_{c_i}(\boldsymbol{z}_i)^\top - \nabla\mathcal{T}_k(\boldsymbol{z}_i)^\top\big)\,\boldsymbol{h}_i.\end{aligned}$$

Definition 5 (Clean pairwise margin)**.** *For every competing class $k \neq c_i$, we define the clean pairwise margin as*

$$\Delta_{ik} = \mathcal{T}_{c_i}(\boldsymbol{z}_i) - \mathcal{T}_k(\boldsymbol{z}_i).$$

Definition 6 (Pairwise margin gradient)**.** *For every competing class $k \neq c_i$, the pairwise margin gradient $\boldsymbol{a}_{ik} \in \mathbb{R}^m$ is defined as*

$$\boldsymbol{a}_{ik} = \nabla\mathcal{T}_{c_i}(\boldsymbol{z}_i) - \nabla\mathcal{T}_k(\boldsymbol{z}_i) = \boldsymbol{J}_i^\top(\boldsymbol{e}_{c_i} - \boldsymbol{e}_k),$$

where $\boldsymbol{e}_k$ denotes the k-th canonical basis vector of $\mathbb{R}^K$. Equivalently, $\boldsymbol{a}_{ik}^\top$ is obtained by subtracting the k-th row of $\boldsymbol{J}_i$ from its c_i-th row.

Definition 7 (Deterministic first-order margin)**.** *Substituting the perturbation $\boldsymbol{h}_i = \boldsymbol{d}_i + \boldsymbol{L}\boldsymbol{\xi}$ into the first-order approximation of the pairwise score gap yields*

$$\begin{aligned}\mathcal{T}_{c_i}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}_k(\boldsymbol{z}_i + \boldsymbol{h}_i) &\approx (\mathcal{T}_{c_i}(\boldsymbol{z}_i) - \mathcal{T}_k(\boldsymbol{z}_i)) + \left(\nabla\mathcal{T}_{c_i}(\boldsymbol{z}_i)^\top - \nabla\mathcal{T}_k(\boldsymbol{z}_i)^\top\right)\boldsymbol{h}_i \\ &= \Delta_{ik} + \boldsymbol{a}_{ik}^\top \boldsymbol{d}_i + \boldsymbol{a}_{ik}^\top \boldsymbol{L}\boldsymbol{\xi}.\end{aligned}$$

Isolating the non-random components, we define the deterministic first-order margin after clipping as

$$\bar{\Delta}_{ik} = \Delta_{ik} + \boldsymbol{a}_{ik}^\top \boldsymbol{d}_i.$$

Definition 8 (Noise-induced pairwise perturbation)**.** *Define the noise-induced pairwise perturbation by*

$$\Xi_{ik} = \boldsymbol{a}_{ik}^\top \boldsymbol{L}\boldsymbol{\xi}.$$

Define its variance by

$$\mathrm{Var}(\Xi_{ik}) = \sigma^2 \sum_{j=1}^{m} \lambda_j \left(\boldsymbol{a}_{ik}^\top \boldsymbol{u}_j\right)^2.$$

Define the largest coordinate-wise projected noise scale by

$$M_{ik} = \max_{1\le j\le m} \sqrt{\lambda_j}\left|\boldsymbol{a}_{ik}^\top \boldsymbol{u}_j\right|,$$

Proposition 6 (Prediction disagreement bound for nonlinear ReLU classifiers)**.** *Suppose that $\mathcal{T} : \mathbb{R}^m \to \mathbb{R}^K$ is a piecewise-affine ReLU network and is differentiable at $\boldsymbol{z}_i$. Let $\mathcal{R}_i$ be the activation cell containing $\boldsymbol{z}_i$ on which $\mathcal{T}$ is affine with Jacobian $\boldsymbol{J}_i$, and let $\mathcal{A}_i = \{\boldsymbol{z}_i + t\boldsymbol{h}_i \in \mathcal{R}_i \text{ for all } t \in [0,1]\}$ denote the event that the perturbation remains in the same activation cell.*

For each competing class $k \neq c_i$, suppose that the deterministic first-order margin after clipping satisfies

$$\bar{\Delta}_{ik} = \Delta_{ik} + \boldsymbol{a}_{ik}^\top \boldsymbol{d}_i > 0.$$

For each pair satisfying $\mathrm{Var}(\Xi_{ik}) > 0$, define

$$\phi_{ik} = \frac{\bar{\Delta}_{ik}^2}{4\,\mathrm{Var}(\Xi_{ik})}, \qquad \psi_{ik} = \frac{\bar{\Delta}_{ik}}{2\sigma M_{ik}}.$$

Then the prediction-disagreement probability of the nonlinear classifier satisfies

$$\Pr_{\boldsymbol{\xi}}\left[\mathcal{C}(\hat{\boldsymbol{z}}_i) \neq \mathcal{C}(\boldsymbol{z}_i)\right] \le \Pr_{\boldsymbol{\xi}}(\mathcal{A}_i^c) + \sum_{\substack{k\neq c_i \\ \mathrm{Var}(\Xi_{ik})>0}} \exp\left(-\min\{\phi_{ik}, \psi_{ik}\}\right).$$

If $\mathrm{Var}(\Xi_{ik}) = 0$, then all projected coefficients vanish and hence $\Xi_{ik} \equiv 0$. Since $\bar{\Delta}_{ik} > 0$, the corresponding pair contributes zero probability.

Proof. On the event $\mathcal{A}_i$, the perturbation remains inside the activation cell $\mathcal{R}_i$. Since $\mathcal{T}$ is affine on $\mathcal{R}_i$, the first-order expression is exact:

$$\mathcal{T}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}(\boldsymbol{z}_i) = \boldsymbol{J}_i \boldsymbol{h}_i.$$

Therefore, for every competing class $k \neq c_i$, on $\mathcal{A}_i$ the actual pairwise score gap satisfies

$$\mathcal{T}_{c_i}(\boldsymbol{z}_i + \boldsymbol{h}_i) - \mathcal{T}_k(\boldsymbol{z}_i + \boldsymbol{h}_i) = \Delta_{ik} + \boldsymbol{a}_{ik}^\top \boldsymbol{h}_i = \Delta_{ik} + \boldsymbol{a}_{ik}^\top \boldsymbol{d}_i + \boldsymbol{a}_{ik}^\top \boldsymbol{L}\boldsymbol{\xi} = \bar{\Delta}_{ik} + \Xi_{ik}.$$

If the nonlinear prediction changes while $\mathcal{A}_i$ occurs, then the score of at least one competing class must reach or exceed the score of the clean predicted class. Thus,

$$\{\mathcal{C}(\hat{\boldsymbol{z}}_i) \neq \mathcal{C}(\boldsymbol{z}_i)\} \subseteq \mathcal{A}_i^c \cup \bigcup_{k\neq c_i} \left\{\Xi_{ik} \le -\bar{\Delta}_{ik}\right\}.$$

Applying the union bound gives

$$\Pr_{\boldsymbol{\xi}}\left[\mathcal{C}(\hat{\boldsymbol{z}}_i) \neq \mathcal{C}(\boldsymbol{z}_i)\right] \leq \Pr_{\boldsymbol{\xi}}(\mathcal{A}_i^c) + \sum_{k \neq c_i} \Pr_{\boldsymbol{\xi}}\left[\Xi_{ik} \leq -\bar{\Delta}_{ik}\right].$$

Define $\kappa_{ikj} = \sqrt{\lambda_j}\left(\boldsymbol{a}_{ik}^\top \boldsymbol{u}_j\right)$, so that $\Xi_{ik} = \sum_{j=1}^m \kappa_{ikj}\xi_j$ and $M_{ik} = \max_j |\kappa_{ikj}|$.

For $\xi_j \sim \mathrm{Lap}(0, \sigma/\sqrt{2})$,

$$\mathbb{E}[e^{t\xi_j}] = \left(1 - \frac{\sigma^2 t^2}{2}\right)^{-1}, \qquad |t| < \frac{\sqrt{2}}{\sigma}.$$

By independence,

$$\log \mathbb{E}\left[e^{t\Xi_{ik}}\right] = \sum_{j=1}^m -\log\left(1 - \frac{\sigma^2 \kappa_{ikj}^2 t^2}{2}\right).$$

For $|t| \leq t_{\max} = \frac{1}{\sigma M_{ik}}$, we have $\sigma^2\kappa_{ikj}^2 t^2/2 \leq 1/2$ for every j. Using $-\log(1-x) \leq 2x$ for $x \in [0, 1/2]$ yields

$$\log \mathbb{E}\left[e^{t\Xi_{ik}}\right] \leq \sigma^2 t^2 \sum_{j=1}^m \kappa_{ikj}^2 = t^2 \operatorname{Var}(\Xi_{ik}).$$

Thus, for any $\tau > 0$ and $0 < t \leq t_{\max}$, the Chernoff bound gives

$$\Pr[\Xi_{ik} \geq \tau] \leq \exp\left(-t\tau + t^2 \operatorname{Var}(\Xi_{ik})\right).$$

If $\tau \leq \frac{2\operatorname{Var}(\Xi_{ik})}{\sigma M_{ik}}$, the unconstrained minimizer $t^\star = \frac{\tau}{2\operatorname{Var}(\Xi_{ik})}$ is admissible, which gives

$$\Pr[\Xi_{ik} \geq \tau] \leq \exp\left(-\frac{\tau^2}{4\operatorname{Var}(\Xi_{ik})}\right).$$

Otherwise, taking $t = t_{\max}$ gives

$$-t_{\max}\tau + t_{\max}^2 \operatorname{Var}(\Xi_{ik}) = -\frac{\tau}{\sigma M_{ik}} + \frac{\operatorname{Var}(\Xi_{ik})}{\sigma^2 M_{ik}^2} < -\frac{\tau}{2\sigma M_{ik}},$$

and

$$\Pr[\Xi_{ik} \geq \tau] \leq \exp\left(-\frac{\tau}{2\sigma M_{ik}}\right).$$

Combining the two regimes,

$$\Pr[\Xi_{ik} \geq \tau] \leq \exp\left[-\min\left\{\frac{\tau^2}{4\operatorname{Var}(\Xi_{ik})}, \frac{\tau}{2\sigma M_{ik}}\right\}\right].$$

Since Ξ_{ik} is symmetric about zero, the same bound holds for the left tail:

$$\Pr[\Xi_{ik} \leq -\tau] \leq \exp\left[-\min\left\{\frac{\tau^2}{4\operatorname{Var}(\Xi_{ik})}, \frac{\tau}{2\sigma M_{ik}}\right\}\right].$$

Substituting $\tau = \bar{\Delta}_{ik} > 0$ gives

$$\Pr\left[\Xi_{ik} \leq -\bar{\Delta}_{ik}\right] \leq \exp\left(-\min\{\phi_{ik}, \psi_{ik}\}\right).$$

For $\operatorname{Var}(\Xi_{ik}) = 0$, all projected coefficients vanish, and hence $\Xi_{ik} \equiv 0$. Since $\bar{\Delta}_{ik} > 0$, it follows that

$$\Pr\left[\Xi_{ik} \leq -\bar{\Delta}_{ik}\right] = 0.$$

Combining the positive-variance tail bounds with the zero-variance case and applying the union bound above proves the result. □

Remark 2. *The proposition separates nonlinear prediction disagreement into two sources: activation-region crossing and within-region pairwise margin flipping. The latter is controlled by the projected noise variance and the largest coordinate-wise projected noise scale. We use this result to characterize prediction stability of the proposed transformation; we do not claim that the resulting disagreement bound is jointly optimized with respect to clipping and noise allocation.*

A.5 Sensitivity Amplification in Transformed Space

Let $\boldsymbol{\phi} = \boldsymbol{z} - \boldsymbol{z}'$ be a perturbation in the original space $\mathcal{Z}$, and $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^\top$ be the covariance matrix. The worst-case sensitivity of the transformed space under the inverse Cholesky factor $\boldsymbol{L}^{-1}$, defined by the induced matrix norm $\|\boldsymbol{L}^{-1}\|_p = \max_{\boldsymbol{\phi}\neq\mathbf{0}} \frac{\|\boldsymbol{L}^{-1}\boldsymbol{\phi}\|_p}{\|\boldsymbol{\phi}\|_p}$, is amplified ($\|\boldsymbol{L}^{-1}\|_p > 1$) if the transformation reduces variance along any principal coordinate, for both $p \in \{1, 2\}$.

Proof. The maximum perturbation bound is fundamentally determined by the induced ℓ_p-norm of the transformation matrix: $\|\boldsymbol{L}^{-1}\boldsymbol{\phi}\|_p \leq \|\boldsymbol{L}^{-1}\|_p \|\boldsymbol{\phi}\|_p$. We prove the sensitivity amplification for each norm independently.

Case 1: ℓ_2-norm ($p = 2$). The induced ℓ_2-norm (spectral norm) of a matrix is its maximum singular value. For $\boldsymbol{L}^{-1}$, this is related to the eigenvalues of the precision matrix $\boldsymbol{\Sigma}^{-1}$:

$$\|\boldsymbol{L}^{-1}\|_2 = \sqrt{\lambda_{\max}((\boldsymbol{L}^{-1})^\top \boldsymbol{L}^{-1})} = \sqrt{\lambda_{\max}(\boldsymbol{\Sigma}^{-1})} = \frac{1}{\sqrt{\lambda_{\min}(\boldsymbol{\Sigma})}}$$

If $\boldsymbol{L}$ applies a pure rotation without rescaling ($\boldsymbol{\Sigma} = \boldsymbol{I}$), then $\lambda_{\min}(\boldsymbol{\Sigma}) = 1$, yielding $\|\boldsymbol{L}^{-1}\|_2 = 1$ (distance is perfectly preserved). However, if the transformation shrinks the space such that the minimum variance is less than 1 ($\lambda_{\min}(\boldsymbol{\Sigma}) < 1$), the bound becomes strictly greater than 1:

$$\|\boldsymbol{L}^{-1}\|_2 > 1 \implies \max_{\boldsymbol{\phi}\neq\mathbf{0}} \frac{\|\boldsymbol{L}^{-1}\boldsymbol{\phi}\|_2}{\|\boldsymbol{\phi}\|_2} > 1$$

Thus, the worst-case ℓ_2 sensitivity increases under variance reduction.

Case 2: ℓ_1-norm ($p = 1$). The induced ℓ_1-norm of a matrix is its maximum absolute column sum: $\|\boldsymbol{L}^{-1}\|_1 = \max_j \sum_i |(\boldsymbol{L}^{-1})_{ij}|$. Unlike the ℓ_2-norm, the ℓ_1-norm is coordinate-dependent. For a pure rotation matrix $\boldsymbol{U}$ (where no coordinate scaling occurs), the rotation misaligns the canonical axes, generally resulting in $\|\boldsymbol{U}^{-1}\|_1 \geq 1$. Furthermore, if $\boldsymbol{L}$ involves shrinking (rescaling), we utilize the submultiplicative property of induced norms on the identity matrix $\boldsymbol{I} = \boldsymbol{L}\boldsymbol{L}^{-1}$:

$$1 = \|\boldsymbol{I}\|_1 \leq \|\boldsymbol{L}\|_1 \|\boldsymbol{L}^{-1}\|_1 \implies \|\boldsymbol{L}^{-1}\|_1 \geq \frac{1}{\|\boldsymbol{L}\|_1}$$

If the transformed space is compressed such that the absolute column sums of $\boldsymbol{L}$ are bounded below 1 ($\|\boldsymbol{L}\|_1 < 1$), it guarantees that the inverse transformation expands the space:

$$\|\boldsymbol{L}^{-1}\|_1 \geq \frac{1}{\|\boldsymbol{L}\|_1} > 1$$

Combining the effects of rotation and scaling, if the transformation satisfies $\|L\|_1 < 1$, then $\|L^{-1}\|_1 > 1$, thereby amplifying the worst-case ℓ_1 sensitivity.

A.6 Privacy Guarantee

Neighboring relation: Under the standard (ϵ, δ)-LDP definition, the privacy inequality must hold for *every* pair of possible local inputs $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$. Expressed as a neighboring relation, this is the complete relation $\mathcal{Z} \times \mathcal{Z}$: an arbitrary replacement of the entire representation.

Theorem 1 (Privacy Guarantee of Jacobian-Guided Anisotropic Noise Reshaping)**.** *Let $\mathcal{Z} \subseteq \mathbb{R}^m$ be the domain of a single private representation, which may possibly be unbounded. Assume:*

- ***(A1) Public configuration.*** *Conditional on fixed public side information, let $\boldsymbol{U}$ be orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ satisfy $0 < \lambda_{\min} \leq \lambda_i \leq \lambda_{\max} < \infty$. Define the configuration tuple $\boldsymbol{\phi} = (\boldsymbol{U}, \boldsymbol{\Lambda}, \boldsymbol{\mu})$, and let $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}$ (which is invertible and has a finite condition number). Let $\rho > 0$ and $p \in [1, \infty]$. $\boldsymbol{\phi}$, p, and ρ may be chosen analytically or empirically using only the public task model, public data, and public validation results. Once selected, they are fixed before any input is privatized, and are neither computed from nor adapted to the protected representation.*
- ***(A2) Bounded intermediate domain.*** *With $\theta_p(\bar{\boldsymbol{z}}; \rho) = \min(1, \rho/\|\bar{\boldsymbol{z}}\|_p)$ for $\bar{\boldsymbol{z}} \neq \mathbf{0}$ and 1 otherwise, define $f_\phi(\boldsymbol{z}) = \theta_p\big(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}); \rho\big)\, \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu})$ and the bounded domain $\bar{\mathcal{Z}}_{p,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_p \leq \rho\}$. Radial clipping ensures $f_{\boldsymbol{\phi}}(\mathcal{Z}) \subseteq \bar{\mathcal{Z}}_{p,\rho}$ for every $\mathcal{Z}$, regardless of whether it is bounded or not.*

- ***(A3) Underlying randomizer.*** *The base randomizer $\mathcal{M}$ is instantiated with the domain and calibration required by its original privacy theorem. Under the corresponding conditions, $\mathcal{M}$ satisfies (ϵ, δ)-LDP uniformly over $\bar{\mathcal{Z}}_{p,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_p \le \rho\}$.*
- ***(A4) Post-processing.*** *$g_{\boldsymbol{\phi}}(\boldsymbol{v}) = \boldsymbol{L}\boldsymbol{v} + \boldsymbol{\mu}$ is deterministic, measurable, and fixed solely from public information.*

Then for every admissible configuration $\boldsymbol{\phi}$, the complete mechanism $\Phi_{\boldsymbol{\phi}} = g_{\boldsymbol{\phi}} \circ \mathcal{M} \circ f_{\boldsymbol{\phi}}$ satisfies (ϵ, δ)-LDP on $\mathcal{Z}$.

Proof. Fix arbitrary $\boldsymbol{z}, \boldsymbol{z}' \in \mathcal{Z}$ and a measurable set $S \subseteq \mathbb{R}^m$. The public configuration $\boldsymbol{\phi}$ is fixed under the neighboring comparison; by (A1) it does not depend on the input, so the same deterministic map $f_{\boldsymbol{\phi}}$ is applied to both $\boldsymbol{z}$ and $\boldsymbol{z}'$. By (A2),

$$f_{\boldsymbol{\phi}}(\boldsymbol{z}), f_{\boldsymbol{\phi}}(\boldsymbol{z}') \in \bar{\mathcal{Z}}_{p,\rho}.$$

Therefore, by (A3), the uniform (ϵ, δ)-LDP guarantee of $\mathcal{M}$ applies to this pair.

Since $g_{\boldsymbol{\phi}}$ is deterministic and measurable, its pre-image

$$T = g_{\boldsymbol{\phi}}^{-1}(S) = \{\boldsymbol{v} \in \mathbb{R}^m : g_{\boldsymbol{\phi}}(\boldsymbol{v}) \in S\}$$

is measurable. Hence,

$$\Pr[\Phi_{\boldsymbol{\phi}}(\boldsymbol{z}) \in S] = \Pr[\mathcal{M}(f_{\boldsymbol{\phi}}(\boldsymbol{z})) \in T] \le e^{\epsilon} \Pr[\mathcal{M}(f_{\boldsymbol{\phi}}(\boldsymbol{z}')) \in T] + \delta = e^{\epsilon} \Pr[\Phi_{\boldsymbol{\phi}}(\boldsymbol{z}') \in S] + \delta.$$

Since $\boldsymbol{z}$, $\boldsymbol{z}'$, and S were arbitrary, $\Phi_{\boldsymbol{\phi}}$ satisfies (ϵ, δ)-LDP on $\mathcal{Z}$. The pure ϵ-LDP result follows as the special case $\delta = 0$. □

We provide the instantiations of randomizers, including Laplace, AGM, and PrivUnit, as Corollaries below.

Corollary 4 (Laplace mechanism (Dwork and Roth, 2014))**.** *Let $p = 1$, so that the intermediate domain is*

$$\bar{\mathcal{Z}}_{1,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_1 \le \rho\},$$

and let the output space be $\mathcal{Y} = \mathbb{R}^m$. Because the LDP neighboring relation is the complete relation, the relevant sensitivity is the diameter of the entire intermediate domain. For every $\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{1,\rho}$,

$$\|\bar{\boldsymbol{z}} - \bar{\boldsymbol{z}}'\|_1 \le \|\bar{\boldsymbol{z}}\|_1 + \|\bar{\boldsymbol{z}}'\|_1 \le 2\rho.$$

This bound is attained by $\bar{\boldsymbol{z}} = \rho\boldsymbol{e}_1$ and $\bar{\boldsymbol{z}}' = -\rho\boldsymbol{e}_1$. Hence,

$$\Delta_1 = \sup_{\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{1,\rho}} \|\bar{\boldsymbol{z}} - \bar{\boldsymbol{z}}'\|_1 = 2\rho.$$

Define the coordinate-wise Laplace randomizer

$$\mathcal{M}_{\mathrm{Lap}}(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{z}} + \boldsymbol{\xi}, \qquad \xi_j \overset{\text{i.i.d.}}{\sim} \mathrm{Lap}(0, b), \qquad b = \frac{2\rho}{\epsilon}.$$

For every output point $\boldsymbol{y} \in \mathbb{R}^m$ and every $\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{1,\rho}$, the ratio of the corresponding densities satisfies

$$\frac{q(\boldsymbol{y} \mid \bar{\boldsymbol{z}})}{q(\boldsymbol{y} \mid \bar{\boldsymbol{z}}')} = \exp\left(\frac{\|\boldsymbol{y} - \bar{\boldsymbol{z}}'\|_1 - \|\boldsymbol{y} - \bar{\boldsymbol{z}}\|_1}{b}\right) \le \exp\left(\frac{\|\bar{\boldsymbol{z}} - \bar{\boldsymbol{z}}'\|_1}{b}\right) \le \exp\left(\frac{2\rho}{b}\right) = e^{\epsilon}.$$

Integrating this pointwise inequality over any measurable set $T \subseteq \mathbb{R}^m$ gives

$$\Pr[\mathcal{M}_{\mathrm{Lap}}(\bar{\boldsymbol{z}}) \in T] \le e^{\epsilon} \Pr[\mathcal{M}_{\mathrm{Lap}}(\bar{\boldsymbol{z}}') \in T].$$

Therefore, $\mathcal{M}_{\mathrm{Lap}}$ satisfies Assumption (A3) with $\delta = 0$ uniformly over $\bar{\mathcal{Z}}_{1,\rho}$. The input-domain map is

$$f_{\boldsymbol{\phi}}(\boldsymbol{z}) = \theta_1\left(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}); \rho\right) \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}),$$

and the public reconstruction map is

$$g_{\boldsymbol{\phi}}(\boldsymbol{v}) = \boldsymbol{L}\boldsymbol{v} + \boldsymbol{\mu}.$$

No separate zero-vector convention is required, because the Laplace randomizer is defined on every point of $\bar{\mathcal{Z}}_{1,\rho}$, including the origin. Consequently,

$$\Phi_{\mathrm{Lap},\boldsymbol{\phi}} = g_{\boldsymbol{\phi}} \circ \mathcal{M}_{\mathrm{Lap}} \circ f_{\boldsymbol{\phi}}$$

satisfies pure ϵ-LDP on the original representation domain $\mathcal{Z}$. The calibration $b = 2\rho/\epsilon$ depends only on the chosen norm, radius, and privacy budget; neither $\boldsymbol{U}$ nor $\boldsymbol{\Lambda}$ enters the calibration. Thus, the public anisotropic rotation and rescaling do not incur an additional privacy cost.

Corollary 5 (Analytically calibrated Gaussian mechanism (Balle and Wang, 2018))**.** *Let $p = 2$, so that the intermediate domain is*

$$\bar{\mathcal{Z}}_{2,\rho} = \{\bar{\boldsymbol{z}} \in \mathbb{R}^m : \|\bar{\boldsymbol{z}}\|_2 \le \rho\},$$

and let the output space be $\mathcal{Y} = \mathbb{R}^m$. Because the LDP neighboring relation is the complete relation, the relevant ℓ_2-sensitivity is the diameter of the entire intermediate domain. For every $\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{2,\rho}$,

$$\|\bar{\boldsymbol{z}} - \bar{\boldsymbol{z}}'\|_2 \le \|\bar{\boldsymbol{z}}\|_2 + \|\bar{\boldsymbol{z}}'\|_2 \le 2\rho.$$

This bound is attained by $\bar{\boldsymbol{z}} = \rho \boldsymbol{e}_1$ and $\bar{\boldsymbol{z}}' = -\rho \boldsymbol{e}_1$. Hence,

$$\Delta_2 = \sup_{\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{2,\rho}} \|\bar{\boldsymbol{z}} - \bar{\boldsymbol{z}}'\|_2 = 2\rho.$$

Define the isotropic Gaussian randomizer

$$\mathcal{M}_{\mathrm{G}}(\bar{\boldsymbol{z}}) = \bar{\boldsymbol{z}} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_m\right).$$

For prescribed $\epsilon > 0$ and $\delta \in (0, 1)$, choose $\sigma > 0$ according to the analytical Gaussian calibration for sensitivity $\Delta_2 = 2\rho$; equivalently, choose σ such that

$$\Psi\left(\frac{\Delta_2}{2\sigma} - \frac{\epsilon\sigma}{\Delta_2}\right) - e^{\epsilon}\Psi\left(-\frac{\Delta_2}{2\sigma} - \frac{\epsilon\sigma}{\Delta_2}\right) \le \delta,$$

where Ψ denotes the cumulative distribution function of the standard normal distribution. Under this calibration, for every $\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{2,\rho}$ and every measurable set $T \subseteq \mathbb{R}^m$,

$$\Pr[\mathcal{M}_{\mathrm{AGM}}(\bar{\boldsymbol{z}}) \in T] \le e^{\epsilon} \Pr[\mathcal{M}_{\mathrm{AGM}}(\bar{\boldsymbol{z}}') \in T] + \delta.$$

Therefore, $\mathcal{M}_{\mathrm{AGM}}$ satisfies Assumption (A3) uniformly over $\bar{\mathcal{Z}}_{2,\rho}$. The input-domain map is

$$f_{\boldsymbol{\phi}}(\boldsymbol{z}) = \theta_2\left(\boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}); \rho\right) \boldsymbol{L}^{-1}(\boldsymbol{z} - \boldsymbol{\mu}),$$

and the public reconstruction map is

$$g_{\boldsymbol{\phi}}(\boldsymbol{v}) = \boldsymbol{L}\boldsymbol{v} + \boldsymbol{\mu}.$$

No separate zero-vector convention is required, because the Gaussian randomizer is defined on every point of $\bar{\mathcal{Z}}_{2,\rho}$, including the origin. Consequently,

$$\Phi_{\mathrm{AGM},\boldsymbol{\phi}} = g_{\boldsymbol{\phi}} \circ \mathcal{M}_{\mathrm{AGM}} \circ f_{\boldsymbol{\phi}}$$

satisfies (ϵ, δ)-LDP on the original representation domain $\mathcal{Z}$. The analytical calibration depends on (ϵ, δ, ρ) through the sensitivity $\Delta_2 = 2\rho$; neither the public basis $\boldsymbol{U}$ nor the finite scaling spectrum $\boldsymbol{\Lambda}$ enters the calibration. Thus, the public anisotropic rotation and rescaling do not incur an additional privacy cost. This instantiation is included to show that the proposed wrapper also composes with approximate-LDP mechanisms.

Corollary 6 (PrivUnit mechanism (Bhowmick et al., 2018; Asi et al., 2022))**.** *Let the output space be $\mathcal{Y} = \mathbb{S}^{m-1}$. PrivUnit2 (Bhowmick et al., 2018) satisfies pure ϵ-LDP uniformly over $\mathbb{S}^{m-1}$. Define*

$$N(\bar{\boldsymbol{z}}) = \begin{cases} \bar{\boldsymbol{z}}/\|\bar{\boldsymbol{z}}\|_2, & \bar{\boldsymbol{z}} \neq \boldsymbol{0}, \\ \boldsymbol{\eta}_0, & \bar{\boldsymbol{z}} = \boldsymbol{0}, \end{cases}$$

where $\boldsymbol{\eta}_0 \in \mathbb{S}^{m-1}$ is a fixed public unit vector, and let

$$\mathcal{M}_{\mathrm{PU2}} = \mathrm{PU2}_{\epsilon,m} \circ N.$$

For any $\bar{\boldsymbol{z}}, \bar{\boldsymbol{z}}' \in \bar{\mathcal{Z}}_{p,\rho}$, their normalized images lie in $\mathbb{S}^{m-1}$; hence the sphere-level guarantee applies directly, and $\mathcal{M}_{\mathrm{PU2}}$ satisfies Assumption (A3) with $\delta = 0$. Moreover, for every non-zero $\bar{\boldsymbol{z}}$,

$$N\big(\mathrm{Clip}_p(\bar{\boldsymbol{z}}; \rho)\big) = N(\bar{\boldsymbol{z}}),$$

because radial clipping multiplies its input by a positive scalar. Thus, p and ρ do not affect the PrivUnit2 input direction. Let $\tau_{\epsilon,m} \neq 0$ be the publicly fixed cap threshold selected by the optimized implementation as a fixed point of its mean directional response; in the implementation considered here, the cap threshold and directional debiasing constant therefore coincide. The reconstruction

$$g_{\mathrm{PU2},\boldsymbol{\phi}}(\boldsymbol{y}) = \boldsymbol{L}\left(\frac{\rho}{\tau_{\epsilon,m}}\boldsymbol{y}\right) + \boldsymbol{\mu}$$

uses only public constants. Therefore, no input-specific transformed norm is released and no additional privacy budget is consumed. Consequently,

$$\Phi_{\mathrm{PU2},\boldsymbol{\phi}} = g_{\mathrm{PU2},\boldsymbol{\phi}} \circ \mathcal{M}_{\mathrm{PU2}} \circ f_{\boldsymbol{\phi}}$$

satisfies pure ϵ-LDP on $\mathcal{Z}$. The same preprocessing and composition argument applies to PrivUnitG (Asi et al., 2022) under its cited pure ϵ-LDP guarantee for unit-sphere inputs, with its Gaussian-based $\mathbb{R}^m$-valued randomizer and mechanism-specific public reconstruction map.

Corollary 7 (Invariance of the certified privacy parameters)**.** *Fix the base randomizer and its mechanism-specific privacy calibration, together with any public quantities on which that calibration depends. For the additive mechanisms, these quantities include the clipping norm p and radius ρ. Let $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ be any two publicly selected configurations satisfying (A1)-(A4). Then the corresponding complete mechanisms $\Phi_{\boldsymbol{\phi}_1}$ and $\Phi_{\boldsymbol{\phi}_2}$ both satisfy the same certified (ϵ, δ)-LDP parameters. Consequently, any public choice of the Jacobian basis, active rank, offset, or bounded scaling spectrum may affect utility, but it does not alter the stated privacy parameters, provided that the resulting configuration remains fixed and admissible under the corresponding mechanism-specific assumptions.*

Proof. The theorem applies to each $\boldsymbol{\phi}_i$ separately, and its conclusion depends on $\boldsymbol{\phi}_i$ only through the admissibility conditions (A1)-(A4). For PrivUnit-based mechanisms, Corollary 6 gives a stronger specialization: the privacy calibration depends only on (ϵ, m). Thus, the public clipping norm p and radius ρ may also differ across admissible configurations without changing the certified privacy parameters. □

Corollary 8 (Image-level LDP under client-side feature extraction)**.** *Let $\mathcal{X}$ be the domain of private images and let*

$$\mathrm{Enc}_{\mathrm{pub}} : \mathcal{X} \to \mathcal{Z}$$

be a fixed, deterministic, and measurable feature extractor satisfying

$$\mathrm{Enc}_{\mathrm{pub}}(\mathcal{X}) \subseteq \mathcal{Z}.$$

Assume that $\mathrm{Enc}_{\mathrm{pub}}$ is fixed before any private input is processed, executed locally on the client, and that neither the raw image nor the unprivatized representation is released before local randomization. This is the deployment protocol illustrated in Figure 5. For any admissible public configuration $\boldsymbol{\phi}$, let

$$\Phi_{\boldsymbol{\phi}} : \mathcal{Z} \to \mathbb{R}^m$$

be the representation-level mechanism from the preceding theorem, satisfying (ϵ, δ)-LDP on $\mathcal{Z}$ under the complete neighboring relation $\mathcal{Z} \times \mathcal{Z}$. Define

$$\widetilde{\Phi}_{\boldsymbol{\phi}} = \Phi_{\boldsymbol{\phi}} \circ \mathrm{Enc}_{\mathrm{pub}} : \mathcal{X} \to \mathbb{R}^m.$$

Then $\widetilde{\Phi}_{\boldsymbol{\phi}}$ satisfies (ϵ, δ)-LDP on $\mathcal{X}$ under the complete neighboring relation $\mathcal{X} \times \mathcal{X}$. That is, for every $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and every measurable set $S \subseteq \mathbb{R}^m$,

$$\Pr[\widetilde{\Phi}_{\boldsymbol{\phi}}(\boldsymbol{x}) \in S] \leq e^{\epsilon} \Pr[\widetilde{\Phi}_{\boldsymbol{\phi}}(\boldsymbol{x}') \in S] + \delta.$$

Proof. Fix arbitrary $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ and a measurable set $S \subseteq \mathbb{R}^m$. Since

$$\mathrm{Enc}_{\mathrm{pub}}(\boldsymbol{x}), \mathrm{Enc}_{\mathrm{pub}}(\boldsymbol{x}') \in \mathcal{Z},$$

the (ϵ, δ)-LDP guarantee of $\Phi_{\boldsymbol{\phi}}$ applies directly to this pair. Therefore,

$$\begin{aligned}
\Pr[\widetilde{\Phi}_{\boldsymbol{\phi}}(\boldsymbol{x}) \in S] &= \Pr[\Phi_{\boldsymbol{\phi}}(\mathrm{Enc}_{\mathrm{pub}}(\boldsymbol{x})) \in S] \\
&\leq e^{\epsilon} \Pr[\Phi_{\boldsymbol{\phi}}(\mathrm{Enc}_{\mathrm{pub}}(\boldsymbol{x}')) \in S] + \delta \\
&= e^{\epsilon} \Pr[\widetilde{\Phi}_{\boldsymbol{\phi}}(\boldsymbol{x}') \in S] + \delta.
\end{aligned}$$

Because $\boldsymbol{x}$, $\boldsymbol{x}'$, and S were arbitrary, the claim follows. □

The client-side condition is essential to the image-level interpretation. If the raw image is transmitted to a server before feature extraction or local randomization, that server observes the unprotected image, and the corollary does not provide image-level LDP against it. In that deployment, the stated guarantee applies only to the representation released after randomization.

Limitations. The primary privacy unit of the mechanism is the representation. Image-level LDP follows only under the client-side deployment described in Figure 5, where neither the raw image nor the unprivatized representation is released. Our method also relies on a fixed encoder available on the client and a fixed downstream task model, which limit applicability when such public components are unavailable or poorly matched to the private distribution.

A.7 Sensitivity Bounding

We employ a clipping operation to control sensitivity and reduce the noise injected into task-sensitive subspaces. Applying $\boldsymbol{L}^{-1}$ during the pre-processing step stretches the task-sensitive subspaces and shrinks the task-insensitive ones. This stretching amplifies the overall sensitivity in $\mathbb{R}^m$ (see Appendix A.5), necessitating the injection of a large amount of noise to guarantee ϵ-LDP. If $\boldsymbol{L}$ is applied during post-processing without clipping, the task-sensitive subspaces receive the same amount of noise as they would if isotropic noise were directly injected into the original space $\mathcal{Z}$, while the task-insensitive subspaces suffer from excessive noise amplification. However, by regulating the sensitivity through clipping in the transformed space $\boldsymbol{L}^{-1}\mathcal{Z}$, we can effectively reduce the noise injected into the task-sensitive subspaces while satisfying ϵ-LDP. The task-insensitive subspaces also remain protected by ϵ-LDP-compliant noise, although this noise is often excessive.

A.8 Compatibility

Our approach can be seamlessly integrated with mechanisms that achieve randomization by sampling noise from Laplace, Gaussian, or other distributions and adding it to the representation. Furthermore, it is compatible with mechanisms based on probabilistic sampling designed to preserve directional information (Bhowmick et al., 2018; Asi et al., 2022). This is achieved simply by incorporating an additional step to transform $\tilde{\boldsymbol{z}}$ into the required input format of the target mechanism, followed by an inverse transformation prior to applying the post-processing function.

A.9 Aggregation

Because of the heavy randomization in LDP, practical applications frequently adopt an aggregation step to average out the injected noise. In our proposed approach, this aggregation takes place between the randomization and post-processing steps. Since we inject zero-mean, isotropic noise during randomization, the aggregation step remains unbiased and reduces the variance of the injected noise as the number of samples increases. Our post-processing function is applied to this aggregated output, ensuring that the resulting representations perform effectively in the downstream model.

A.10 Experimental Details

We evaluate our method across three downstream tasks, including one time-series regression and two image classification tasks. The evaluation covers a range of privacy budgets ($\epsilon \in [0.5, 10]$), representation dimensionalities ($m \in \{16, 32, 64\}$), and downstream models, such as linear regression (LR), linear classifiers (LC) and multi-layer perceptron (MLP) classifiers.

A.10.1 Global Setup

Experimental Setup. All experiments were conducted on two workstations: one equipped with an Intel Xeon Platinum 8358 CPU, 64 GB of RAM, and a single NVIDIA L40S GPU, and the other with an Intel Core i7-8700 CPU, 32 GB of RAM, and a single NVIDIA GTX 1080 Ti GPU.

Hyperparameters. All clipping thresholds ρ are set to the 90th percentile of the corresponding norms computed over the public training set. The finite scale cap $\lambda_{\max}$ is fixed to 1000 a priori, independently of the private evaluation data. Appendix A.10.10 reports a post-hoc sensitivity analysis showing that utility is largely insensitive to this choice for $\lambda_{\max} \geq 10^2$.

A.10.2 Setup 1: Regression on London Smart Meters Dataset

Dataset. We evaluate our method using a linear regression (LR) model trained on the publicly available London Smart Meters dataset (D., 2019; UK Power Networks). This dataset records the half-hourly electricity consumption of 5,547 London households over approximately 27 months. The dataset provides several aggregated features for each household, including *household_id*, *day*, *energy_count*, *energy_mean*, *energy_median*, *energy_max*, *energy_min*, and *energy_std*.

Downstream Task. The linear regression (LR) predicts the next day's *energy_mean*. We use a m-dimensional vector $\boldsymbol{z}$ as the input, which is made up of *energy_mean* values from the previous m consecutive days. The LR model is trained via Ordinary Least Squares (OLS).

Privacy Model. We employ *user-level* LDP as our privacy constraint, considering that smart meter readings can expose private household activities. The goal of our experiments is to improve data utility under this localized privacy guarantee.

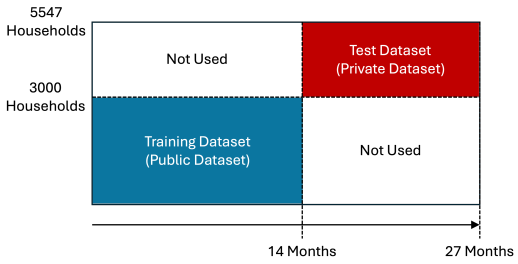


Figure 3: London Smart Meters Dataset Splitting Scheme for Evaluation with Regression

Public and Private Datasets. The dataset is partitioned both by household and chronologically into training and test sets, which serve as the public and private datasets, respectively. Specifically, we utilize the initial 14 months of data from 3,000 randomly selected households as the public dataset to train the LR model. The subsequent 13 months of data from the remaining households constitute the private dataset used to evaluate the mechanisms.

Evaluation Protocol. We adopt an evaluation scenario in which an untrusted server collects and aggregates randomized vectors from households in the private dataset pool before performing the regression task. To avoid privacy budget composition across evaluation rounds, the private household pool of 2,547 households is partitioned into disjoint, equal-sized subsets across the N rounds prior to evaluation. Each round is assigned an exclusive subset of $\lfloor 2547/N \rfloor$ households, ensuring that no household appears in more than one round. On each date, K ($1 \leq K \leq \lfloor 2547/N \rfloor$) households are then uniformly sampled at random without replacement from the round's assigned subset. For instance, with $N = 20$, each round is allocated 127 households, yielding a maximum of $K = 127$ per round. Furthermore, evaluation dates are spaced at least m days apart, ensuring that no household's measurement from a single day appears in more than one evaluation window. Together, these conditions guarantee that each household contributes to at most one aggregation across the entire evaluation, ensuring that the privacy cost incurred is ϵ per household regardless of N.

The evaluation is conducted over N rounds. For instance, setting $m = 16$ corresponds to performing $N = 20$ regression rounds over a span of 320 days.

We evaluate utility by measuring the Mean Squared Error (MSE) against ground-truth values.

Composition. In our evaluation, the sequential composition of privacy budgets is circumvented. As detailed above, the partitioning of the household pool into disjoint subsets and the m-day spacing between evaluation dates ensure that the data from any given household is queried at most once across all N rounds. According to the parallel composition theorem of differential privacy, evaluating queries on disjoint subsets of data does not accumulate the privacy cost. Therefore, the total privacy guarantee for any individual household remains bounded by the single-query budget ϵ, without requiring advanced composition mechanisms or budget degradation.

A.10.3 Setup 2: Image Classification on MNIST and CIFAR-10 Datasets

Dataset. We evaluate our method using linear and nonlinear classifiers pre-trained on two public datasets, MNIST and CIFAR-10: MNIST (Lecun et al., 1998) consists of 70,000 grayscale images of handwritten digits (0–9) with a size of 28×28 pixels. It is partitioned into a training set of 60,000 images and a test set of 10,000 images. CIFAR-10 (Krizhevsky, 2009) is utilized to assess performance on more complex, high-dimensional color images. It contains 60,000 32×32 images across 10 classes, with 6,000 images per class. The dataset is divided into 50,000 training images and 10,000 test images. Additionally, we incorporate CIFAR-10-C (Hendrycks and Dietterich, 2019) to evaluate the robustness of our mechanisms against common real-world corruptions. This dataset consists of the original CIFAR-10 test set subjected to 19 types of corruptions, including numerous forms of noise, blur, weather, and digital effects, across five levels of severity. It serves as a benchmark for assessing the model's performance under distribution shifts and corrupted inputs.

In our experiments, the training sets from CIFAR-10 and MNIST are treated as public. These are partitioned into training and validation sets with an 80:20 split to pre-train the public feature extractors (VAE (Kingma and Welling, 2014) and ResNet-20 (He et al., 2016)) and classifiers. To assess the utility of the mechanisms, the test sets of CIFAR-10-C (Hendrycks and Dietterich, 2019) (e.g., brightness, fog, defocus blur, and Gaussian noise), CIFAR-10, and MNIST are treated as private. Note that the CIFAR-10-C test set exhibits a distribution shift compared to the CIFAR-10 training set.

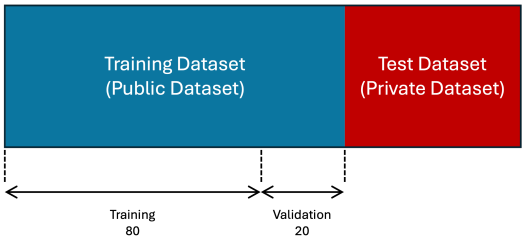


Figure 4: Image Dataset Splitting Scheme for Evaluation with Classification using MNIST and CIFAR-10

Downstream Task. For the classification task, the objective is to predict the class label associated with each input image, such as digits for MNIST or object categories for CIFAR-10 and CIFAR-10-C. We use an m-dimensional representation vector $\boldsymbol{z}$ as the input, which is extracted from the raw image data through a pre-trained feature extractor. The classification models are trained by minimizing the standard cross-entropy loss.

Privacy Model. For the image classification tasks using MNIST, CIFAR-10 and CIFAR-10-C, unlike the regression task, we apply *item-level* LDP, where the privacy guarantee is provided for each individual image sample.

Public and Private Datasets. Consistent with the regression task, the training datasets are treated as public. These are further partitioned into training and validation sets with an 80:20 split to pre-train the public feature extractor and classifiers. Conversely, the test datasets are treated as private, serving as the evaluation data to assess the utility of our method under LDP constraints.

Evaluation Protocol. We consider a distributed setting where a central server aggregates representations $\boldsymbol{z}$ from K virtual clients. These representations are extracted by processing private test images through a public feature extractor. For an N-round evaluation, the 10,000 samples in the private test dataset are partitioned into N disjoint subsets and randomly assigned to clients, regardless of their class labels. For instance, with $N = 10000$ and $K = 1$, each client is allocated 1 unique test sample. To avoid privacy budget accumulation, we ensure that each representation vector is transmitted only once across the N rounds.

Utility Metrics and Composition. Utility is quantified by classification accuracy. From a privacy perspective, because each representation is transmitted only once, the budget does not accumulate over multiple rounds.

A.10.4 Downstream Models

Linear Regression. For the regression task, we evaluate the utility of our representations using the London smart meter dataset. The input feature is an m-dimensional representation vector $\boldsymbol{z}$, constructed from a sliding window of m consecutive days of energy data. The models are trained via OLS. The objective is to predict an energy metric for the following day, such as the *energy_mean*. The OLS algorithm fits the input data to learn a coefficient vector $\boldsymbol{W} \in \mathbb{R}^m$ and an intercept scalar, minimizing the mean squared error (MSE) between the predicted and actual target.

Linear Classifier. In our experiments, we use a multinomial logistic regression model for CIFAR-10 classification. Specifically, we utilize a linear classifier head corresponding to the final fully-connected layer of the ResNet-20 architecture. This head directly maps the fixed 64-dimensional feature vectors, obtained from the pre-trained feature extractor, to 10 class logits. Under this standard pre-training, the model achieves an accuracy over 90% on the CIFAR-10 validation set. Since the classifier relies on a single linear layer mapping to 10 output classes, the Jacobian matrix of its outputs with respect

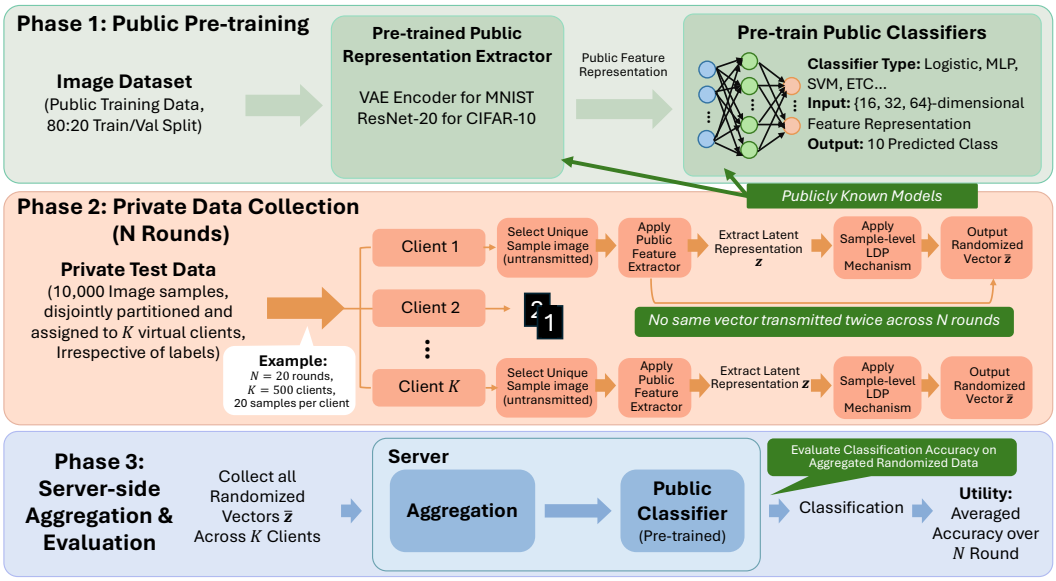


Figure 5: Evaluation Protocol for Classification.

to the input representations has a rank of at most 10. This restricted rank bounds the task-relevant subspace.

Nonlinear MLP Classifier. To demonstrate that our approach generalizes to complex nonlinear settings, we additionally employ Multi-Layer Perceptron (MLP) classifiers. For the CIFAR-10 dataset, the MLP processes 64-dimensional ResNet features through two hidden layers with widths of 10 and 32, interleaved with nonlinear activations (ReLU, GELU, or Tanh). To evaluate our method across different input dimensions, we also apply a ReLU-based MLP with an identical architecture to MNIST classification, taking an m-dimensional latent vector ($m = 32$) from a jointly fine-tuned VAE encoder as input.

A.10.5 Feature Extraction

For the image classification tasks, we employ two distinct models for feature extraction: a standard Variational Autoencoder (VAE) and a ResNet-20 architecture.

The VAE is applied to the MNIST dataset to facilitate experiments across various latent dimensions, enabling us to evaluate the impact of dimensionality on LDP utility. Specifically, we investigate latent representation vectors of size $m = 32$.

Conversely, we utilize ResNet-20 for the CIFAR-10 and CIFAR-10-C datasets. Unlike the simple grayscale digits in MNIST, CIFAR-10 and CIFAR-10-C consist of complex color images, necessitating a feature extractor capable of processing multi-channel inputs. ResNet-20 was chosen for its optimal balance of efficiency and performance; it is a relatively lightweight model that minimizes computational overhead while achieving approximately 90% accuracy on CIFAR-10. To construct the base representations for our evaluations, we utilize the pre-trained ResNet-20 to extract a fixed 64-dimensional feature vector, which serves as the input to the downstream linear classifier.

VAE. For the MNIST dataset, we pre-train a standard Variational Autoencoder (VAE) composed of multi-layer perceptrons (MLPs) using the public dataset. The encoder network takes the flattened 28×28 grayscale images as input and processes them through two shared hidden layers, each containing 512 units with ReLU activations. This is followed by two parallel linear layers that output the mean and log-variance of the m-dimensional latent space. The decoder mirrors this architecture, mapping the latent vector back to the original image space through two 512-unit hidden layers, culminating in a sigmoid output layer. The VAE is trained to minimize the standard objective function, combining binary cross-entropy for reconstruction loss and Kullback-Leibler (KL) divergence. We utilize the Adam optimizer with a learning rate of 10^{-3} and a batch size of 512. To

prevent overfitting, training incorporates early stopping with a patience of 10 epochs, monitoring the validation loss on the 20% split of the public training data.

Following this unsupervised pre-training, the encoder is jointly fine-tuned end-to-end with the downstream classifiers using cross-entropy loss. This joint optimization ensures that the encoder generates task-specific representations optimized for classification accuracy rather than solely for image reconstruction. Finally, for the downstream LDP evaluations, we deterministically extract the mean vector directly from this fine-tuned encoder to serve as the base representation vector, bypassing the reparameterization trick to ensure stable evaluation.

ResNet-20. For the CIFAR-10 and CIFAR-10-C datasets, we implement a standard ResNet-20 architecture as the feature extractor. The network processes the 3-channel color images through an initial convolutional layer (16 channels), followed by three sequential stages of residual blocks. Each stage consists of three basic blocks, progressively increasing the feature channels from 16 to 32 and finally to 64, with spatial downsampling applied via strided convolutions. Network weights are initialized using Kaiming normal initialization. To extract the representation vector $\boldsymbol{z}$, the final feature map is processed through an adaptive average pooling layer and flattened, yielding a 64-dimensional deterministic feature vector.

Similar to the pre-trained classifiers, the feature extractors are pre-trained exclusively on public datasets, specifically the training sets of MNIST and CIFAR-10. During evaluation, these extractors process inputs from private datasets (the test sets of MNIST, CIFAR-10, and CIFAR-10-C) to produce an m-dimensional vector $\boldsymbol{z}$ as output.

A.10.6 Main Results: Comparison with the Task-Aware Approach

We evaluate the performance of our PA-integrated mechanisms against the Task-Aware Approach proposed by Cheng et al. (Cheng et al., 2022) for MNIST classification. Detailed methodological descriptions are available in the original work, and its implementation can be found in the official GitHub repository: `https://github.com/chengjiangnan/task_aware_privacy`. We use their best performance settings.

Mechanisms with fully integrated PA consistently outperform the Task-Aware approach across all privacy budgets. Notably, the Task-Aware approach performs at a level similar to or below Laplace+PA while a low-dimensional latent space of $m = 3$ was selected to achieve its best performance. Given that the scale of the injected noise grows at a rate of $\mathcal{O}(m)$ with increasing dimensionality, this setting provided an advantageous condition for the Task-Aware approach. The performance gap is substantial when comparing the Task-Aware approach to the PrivUnit2+PA and PrivUnitG+PA mechanisms. For instance, at $\epsilon = 7.5$, Task-Aware achieves an accuracy of only 0.2857, whereas PrivUnit2+PA and PrivUnitG+PA reach 0.7252 and 0.6578, respectively, representing a substantial difference of over 0.3. CW+PA w/o bounding underperforms relative to the Task-Aware approach at all ϵ values. This is primarily because CW is optimized for the DMSE objective.

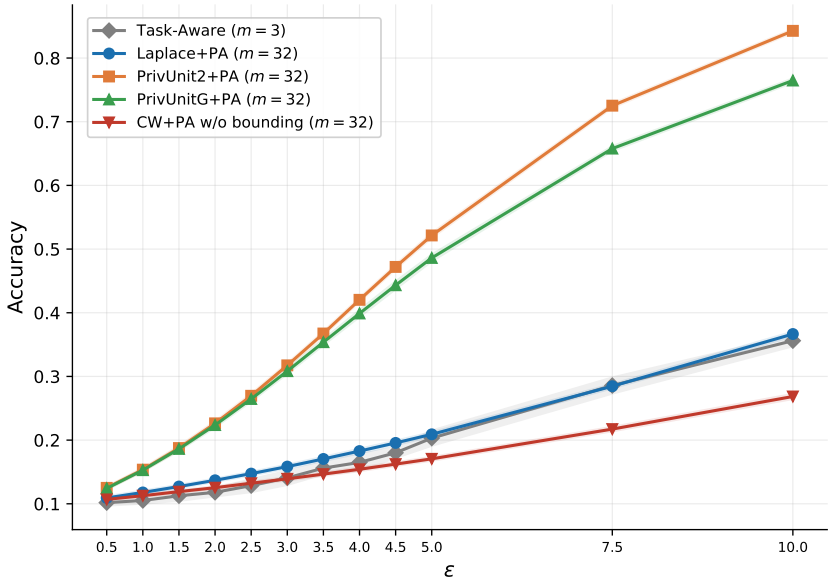


Figure 6: Test accuracy of five mechanisms on MNIST under ϵ-LDP, with shaded regions indicating ± 1 standard deviation. PrivUnit2+PA and PrivUnitG+PA substantially outperform the baselines across all privacy budgets, while Laplace+PA performs comparably to Task-Aware despite using a larger projection dimension ($m = 32$ vs. $m = 3$).

Table 4: Main results for **Nonlinear Classification** on **MNIST** under ϵ-LDP mechanisms. We evaluate utility by measuring the classification accuracy using ground-truth labels. Higher accuracy indicates higher utility.

		Privacy Budget (ϵ)											
Mechanism	m	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Task-Aware	3	0.1018 (0.0049)	0.1052 (0.0053)	0.1126 (0.0083)	0.1180 (0.0067)	0.1285 (0.0105)	0.1407 (0.0114)	0.1560 (0.0101)	0.1650 (0.0135)	0.1801 (0.0084)	0.2031 (0.0123)	0.2857 (0.0130)	0.3561 (0.0091)
Laplace+PA	32	0.1092 (0.0030)	0.1178 (0.0032)	0.1271 (0.0029)	0.1369 (0.0029)	0.1474 (0.0032)	0.1584 (0.0035)	0.1704 (0.0037)	0.1828 (0.0040)	0.1956 (0.0039)	0.2091 (0.0037)	0.2846 (0.0037)	0.3666 (0.0041)
PrivUnit2+PA	32	0.1250 (0.0025)	0.1536 (0.0032)	0.1875 (0.0038)	0.2263 (0.0038)	0.2696 (0.0040)	0.3174 (0.0036)	0.3674 (0.0039)	0.4203 (0.0038)	0.4720 (0.0041)	0.5214 (0.0043)	0.7252 (0.0036)	0.8427 (0.0032)
PrivUnitG+PA	32	0.1242 (0.0027)	0.1528 (0.0025)	0.1862 (0.0035)	0.2236 (0.0040)	0.2646 (0.0044)	0.3083 (0.0047)	0.3538 (0.0051)	0.3989 (0.0059)	0.4432 (0.0066)	0.4861 (0.0057)	0.6578 (0.0042)	0.7647 (0.0037)
CW+PA w/o bounding	32	0.1069 (0.0027)	0.1127 (0.0027)	0.1191 (0.0025)	0.1253 (0.0027)	0.1323 (0.0028)	0.1392 (0.0029)	0.1466 (0.0031)	0.1542 (0.0032)	0.1623 (0.0032)	0.1706 (0.0033)	0.2173 (0.0038)	0.2685 (0.0043)

A.10.7 Main Results: Ablation Study on the Accuracy of Non-Linear MLP Classifiers with ReLU Activation on the CIFAR-10-C (Brightness Corruption At The Highest Severity Level 5) Test Dataset

Clipping Threshold. We investigate the impact of varying clipping thresholds ρ on ReLU-based nonlinear MLP classifiers, evaluated on the CIFAR-10-C dataset under maximum brightness corruption (Level 5). The clipping threshold ρ is determined based on specific percentiles of the public data (e.g., the original CIFAR-10 training set).

In summary, despite the increased distortion of individual representations caused by clipping, the substantial reduction in the injected noise scale outweighs this effect. This suggests that the noise scale, rather than clipping-induced distortion, is the primary bottleneck under LDP constraints. Figure 7 presents the results of integrating PA with the Laplace (Dwork and Roth, 2014), PrivUnit2 (Bhowmick et al., 2018), and PrivUnitG (Asi et al., 2022) mechanisms, across clipping thresholds ρ corresponding to the 10th through 100th percentiles. Overall, the mechanisms integrated with PA achieve higher accuracy when a lower ρ is applied. Tighter clipping thresholds (e.g., at the 10th percentile) consistently yield superior accuracy. Although omitted here for brevity, our approach reached peak performance at a ρ corresponding to the 30th percentile for the ReLU-based MLP classifier operating on VAE-encoded MNIST features.

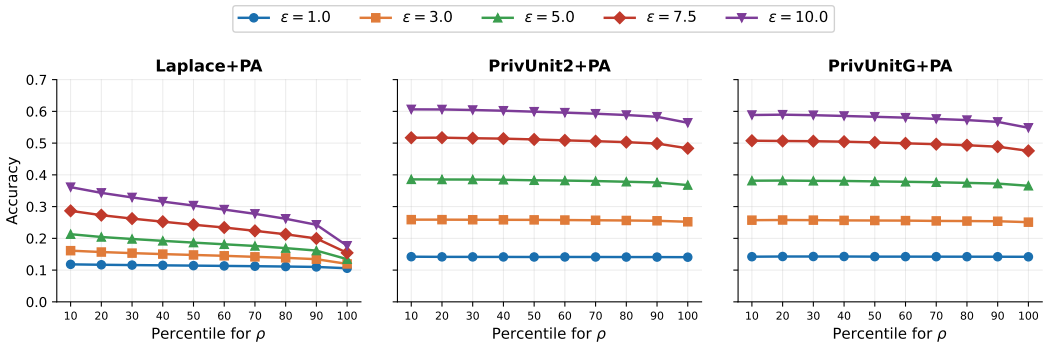


Figure 7: Effect of the clipping threshold ρ on accuracy across privacy budgets $\epsilon \in \{1, 3, 5, 7.5, 10\}$.

Table 5: Ablation study on the **Clipping Threshold** for **Laplace + PA**, **PrivUnit2+PA**, and **PrivUnitG+PA** on **CIFAR-10-C (Brightness Corruption At The Highest Severity Level 5)** Test Dataset under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy

Mechanism	pctile for ρ	Privacy Budget (ϵ) 0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace+PA	10	**0.1083** (0.0018)	**0.1181** (0.0020)	**0.1280** (0.0023)	**0.1385** (0.0026)	**0.1496** (0.0030)	**0.1612** (0.0031)	**0.1735** (0.0032)	**0.1862** (0.0036)	**0.1992** (0.0035)	**0.2131** (0.0037)	**0.2867** (0.0052)	**0.3613** (0.0058)
	20	0.1076 (0.0017)	0.1168 (0.0021)	0.1260 (0.0022)	0.1359 (0.0023)	0.1461 (0.0029)	0.1569 (0.0029)	0.1682 (0.0031)	0.1798 (0.0034)	0.1919 (0.0035)	0.2043 (0.0041)	0.2730 (0.0045)	0.3433 (0.0067)
	30	0.1072 (0.0017)	0.1158 (0.0020)	0.1244 (0.0023)	0.1339 (0.0022)	0.1436 (0.0027)	0.1535 (0.0025)	0.1639 (0.0027)	0.1751 (0.0032)	0.1858 (0.0035)	0.1980 (0.0039)	0.2622 (0.0045)	0.3289 (0.0061)
	40	0.1067 (0.0018)	0.1150 (0.0018)	0.1233 (0.0020)	0.1318 (0.0019)	0.1412 (0.0024)	0.1504 (0.0029)	0.1603 (0.0027)	0.1711 (0.0031)	0.1815 (0.0034)	0.1921 (0.0037)	0.2525 (0.0038)	0.3158 (0.0057)
	50	0.1064 (0.0020)	0.1141 (0.0019)	0.1219 (0.0022)	0.1301 (0.0021)	0.1389 (0.0024)	0.1477 (0.0027)	0.1568 (0.0027)	0.1663 (0.0030)	0.1768 (0.0034)	0.1866 (0.0036)	0.2429 (0.0039)	0.3033 (0.0058)
	60	0.1060 (0.0020)	0.1132 (0.0017)	0.1206 (0.0021)	0.1282 (0.0019)	0.1367 (0.0022)	0.1450 (0.0024)	0.1533 (0.0031)	0.1626 (0.0027)	0.1718 (0.0034)	0.1813 (0.0037)	0.2341 (0.0036)	0.2907 (0.0058)
	70	0.1056 (0.0021)	0.1123 (0.0017)	0.1193 (0.0020)	0.1263 (0.0020)	0.1342 (0.0020)	0.1417 (0.0022)	0.1498 (0.0028)	0.1580 (0.0025)	0.1666 (0.0031)	0.1760 (0.0036)	0.2237 (0.0042)	0.2772 (0.0053)
	80	0.1051 (0.0021)	0.1112 (0.0017)	0.1178 (0.0020)	0.1243 (0.0020)	0.1314 (0.0021)	0.1386 (0.0024)	0.1460 (0.0023)	0.1532 (0.0032)	0.1611 (0.0028)	0.1692 (0.0035)	0.2128 (0.0040)	0.2617 (0.0051)
	90	0.1045 (0.0020)	0.1100 (0.0018)	0.1161 (0.0018)	0.1217 (0.0021)	0.1279 (0.0023)	0.1344 (0.0021)	0.1411 (0.0020)	0.1478 (0.0025)	0.1543 (0.0032)	0.1615 (0.0027)	0.1996 (0.0040)	0.2428 (0.0046)
	100	0.1025 (0.0022)	0.1058 (0.0021)	0.1089 (0.0018)	0.1123 (0.0017)	0.1161 (0.0019)	0.1195 (0.0022)	0.1228 (0.0023)	0.1265 (0.0021)	0.1305 (0.0025)	0.1344 (0.0025)	0.1544 (0.0036)	0.1767 (0.0043)
PrivUnit2+PA	10%	**0.1195** (0.0036)	**0.1422** (0.0040)	**0.1682** (0.0042)	**0.1965** (0.0043)	**0.2274** (0.0037)	**0.2589** (0.0048)	**0.2898** (0.0044)	**0.3235** (0.0047)	**0.3556** (0.0039)	**0.3858** (0.0046)	**0.5169** (0.0049)	**0.6063** (0.0032)
	20%	0.1196 (0.0040)	0.1417 (0.0040)	0.1678 (0.0038)	0.1967 (0.0043)	0.2276 (0.0035)	0.2589 (0.0047)	0.2900 (0.0040)	0.3230 (0.0043)	0.3550 (0.0035)	0.3854 (0.0041)	0.5171 (0.0051)	0.6060 (0.0029)
	30%	0.1196 (0.0040)	0.1416 (0.0038)	0.1674 (0.0033)	0.1965 (0.0045)	0.2273 (0.0042)	0.2586 (0.0047)	0.2901 (0.0039)	0.3229 (0.0047)	0.3548 (0.0033)	0.3852 (0.0036)	0.5154 (0.0044)	0.6043 (0.0033)
	40%	0.1196 (0.0042)	0.1415 (0.0038)	0.1673 (0.0036)	0.1961 (0.0042)	0.2269 (0.0040)	0.2584 (0.0049)	0.2899 (0.0043)	0.3227 (0.0046)	0.3540 (0.0034)	0.3845 (0.0031)	0.5141 (0.0046)	0.6020 (0.0042)
	50%	0.1194 (0.0043)	0.1413 (0.0038)	0.1669 (0.0037)	0.1961 (0.0042)	0.2265 (0.0041)	0.2582 (0.0045)	0.2892 (0.0044)	0.3219 (0.0047)	0.3534 (0.0038)	0.3829 (0.0027)	0.5116 (0.0045)	0.5990 (0.0038)
	60%	0.1196 (0.0043)	0.1414 (0.0038)	0.1667 (0.0035)	0.1956 (0.0043)	0.2262 (0.0040)	0.2577 (0.0048)	0.2886 (0.0043)	0.3209 (0.0050)	0.3522 (0.0042)	0.3820 (0.0025)	0.5087 (0.0046)	0.5960 (0.0045)
	70%	0.1197 (0.0043)	0.1413 (0.0038)	0.1666 (0.0036)	0.1953 (0.0049)	0.2254 (0.0043)	0.2570 (0.0049)	0.2876 (0.0044)	0.3198 (0.0048)	0.3507 (0.0041)	0.3805 (0.0024)	0.5061 (0.0044)	0.5926 (0.0047)
	80%	0.1197 (0.0043)	0.1411 (0.0039)	0.1664 (0.0039)	0.1950 (0.0051)	0.2250 (0.0045)	0.2563 (0.0049)	0.2870 (0.0044)	0.3186 (0.0051)	0.3492 (0.0035)	0.3783 (0.0018)	0.5030 (0.0046)	0.5885 (0.0046)
	90%	0.1196 (0.0043)	0.1410 (0.0039)	0.1662 (0.0037)	0.1945 (0.0050)	0.2244 (0.0048)	0.2553 (0.0053)	0.2856 (0.0048)	0.3171 (0.0051)	0.3471 (0.0036)	0.3760 (0.0023)	0.4986 (0.0046)	0.5829 (0.0047)
	100%	0.1196 (0.0042)	0.1407 (0.0039)	0.1652 (0.0038)	0.1932 (0.0050)	0.2224 (0.0049)	0.2522 (0.0053)	0.2817 (0.0047)	0.3118 (0.0057)	0.3403 (0.0045)	0.3678 (0.0028)	0.4836 (0.0041)	0.5641 (0.0052)
PrivUnitG+PA	10%	**0.1205** (0.0027)	**0.1422** (0.0024)	**0.1677** (0.0031)	**0.1960** (0.0035)	**0.2261** (0.0038)	**0.2573** (0.0043)	**0.2901** (0.0057)	**0.3211** (0.0056)	**0.3516** (0.0053)	**0.3816** (0.0049)	**0.5076** (0.0060)	**0.5885** (0.0057)
	20%	0.1205 (0.0025)	0.1428 (0.0019)	0.1679 (0.0029)	0.1955 (0.0033)	0.2260 (0.0038)	0.2578 (0.0039)	0.2897 (0.0055)	0.3212 (0.0057)	0.3522 (0.0063)	0.3820 (0.0059)	0.5067 (0.0060)	0.5896 (0.0057)
	30%	0.1204 (0.0026)	0.1428 (0.0023)	0.1679 (0.0031)	0.1948 (0.0038)	0.2259 (0.0035)	0.2574 (0.0044)	0.2891 (0.0052)	0.3208 (0.0059)	0.3518 (0.0060)	0.3812 (0.0061)	0.5060 (0.0064)	0.5880 (0.0060)
	40%	0.1203 (0.0028)	0.1428 (0.0024)	0.1678 (0.0030)	0.1950 (0.0035)	0.2256 (0.0034)	0.2566 (0.0038)	0.2887 (0.0050)	0.3201 (0.0057)	0.3513 (0.0061)	0.3808 (0.0061)	0.5045 (0.0065)	0.5855 (0.0068)
	50%	0.1203 (0.0028)	0.1424 (0.0025)	0.1675 (0.0028)	0.1949 (0.0033)	0.2254 (0.0034)	0.2562 (0.0039)	0.2883 (0.0044)	0.3192 (0.0055)	0.3501 (0.0061)	0.3794 (0.0066)	0.5021 (0.0065)	0.5829 (0.0068)
	60%	0.1202 (0.0027)	0.1424 (0.0025)	0.1673 (0.0030)	0.1943 (0.0032)	0.2251 (0.0038)	0.2559 (0.0039)	0.2873 (0.0043)	0.3183 (0.0051)	0.3491 (0.0057)	0.3782 (0.0067)	0.4993 (0.0060)	0.5802 (0.0067)
	70%	0.1205 (0.0027)	0.1423 (0.0026)	0.1671 (0.0032)	0.1940 (0.0031)	0.2242 (0.0034)	0.2549 (0.0042)	0.2867 (0.0040)	0.3176 (0.0049)	0.3476 (0.0060)	0.3767 (0.0064)	0.4966 (0.0060)	0.5762 (0.0067)
	80%	0.1206 (0.0028)	0.1422 (0.0026)	0.1670 (0.0032)	0.1935 (0.0034)	0.2238 (0.0035)	0.2545 (0.0036)	0.2853 (0.0040)	0.3165 (0.0045)	0.3462 (0.0055)	0.3747 (0.0066)	0.4933 (0.0061)	0.5726 (0.0062)
	90%	0.1205 (0.0027)	0.1421 (0.0025)	0.1666 (0.0032)	0.1931 (0.0034)	0.2230 (0.0035)	0.2539 (0.0037)	0.2838 (0.0043)	0.3147 (0.0049)	0.3440 (0.0051)	0.3725 (0.0063)	0.4887 (0.0065)	0.5669 (0.0064)
	100%	0.1207 (0.0027)	0.1419 (0.0026)	0.1660 (0.0034)	0.1922 (0.0034)	0.2210 (0.0036)	0.2510 (0.0038)	0.2804 (0.0046)	0.3094 (0.0049)	0.3382 (0.0051)	0.3656 (0.0048)	0.4756 (0.0064)	0.5483 (0.0068)

A.10.8 Main Results: Accuracy of Nonlinear MLP Classifiers with ReLU Activations on CIFAR-10-C (Brightness, Fog, Defocus Blur, and Gaussian) Test Datasets

A central assumption of our approach is that the Jacobian computed on public data captures the task-critical subspaces of the private data. To stress-test this assumption, we evaluate PA across four corruption types (Brightness, Fog, Defocus Blur, and Gaussian Noise) and five severity levels on CIFAR-10-C, yielding 20 distinct distribution shift scenarios. These corruptions vary not only in severity but also in nature: Brightness and Fog primarily alter global statistics, Defocus Blur degrades spatial frequency, and Gaussian Noise directly corrupts pixel-level information, offering a diverse probe of how different types of public-private mismatch affect the utility of Jacobian-based subspace identification.

A direct comparison of absolute accuracy across severities is confounded by the fact that the no-randomization baseline, which serves as a natural upper bound on achievable utility, itself degrades as corruption severity increases. For instance, under Gaussian Noise at severity 5, the no-randomization accuracy drops to 0.2621, leaving far less room for any mechanism to demonstrate improvement. To account for this, we measure the performance of PA in terms of the fraction of recoverable accuracy it restores. Formally, for each corruption type and severity, we define the recoverable gap as the difference between the no-randomization accuracy and the base mechanism accuracy, and the performance gain as the improvement brought by integrating PA:

$$\text{Recovery Ratio} = \frac{\text{Mean Acc(Baseline+PA)} - \text{Mean Acc(Baseline)}}{\text{Mean Acc(No Randomization)} - \text{Mean Acc(Baseline)}}$$

A higher recovery ratio indicates that PA recovers a larger fraction of the utility lost to LDP noise, independent of the absolute difficulty imposed by the corruption. This metric cleanly isolates PA's contribution from the confounding effect of corruption-induced accuracy degradation.

Tables 6 through 9 present the accuracy of PrivUnit2+PA and PrivUnitG+PA across all 20 distribution shift scenarios, while Figure 8 summarizes the corresponding recovery ratios at representative privacy budgets $\epsilon \in \{1.0, 3.0, 5.0, 7.5\}$.

In summary, the results demonstrate that *PA is robust to both the type and severity of practical distribution shifts*.

These results show several consistent patterns. First, PA consistently enhances utility across all 20 scenarios. The recovery ratio remains positive for every combination of corruption type and severity. This confirms that the Jacobian computed on clean public data remains a viable proxy for identifying task-critical subspaces even when private data undergoes substantial distributional shift. Such findings provide direct empirical support for our central hypothesis.

Second, the recovery ratio scales positively with the privacy budget. Across all corruption types and severities, the ratio grows monotonically with ϵ. This reflects the fact that PA becomes increasingly effective as the noise scale decreases, a trend observed without exception across all 20 scenarios.

However, interpretations of absolute accuracy necessitate careful consideration. The no-randomization accuracy, which serves as the upper bound on recoverable utility, itself degrades as corruption severity increases. This is particularly evident under Gaussian Noise, where the no-randomization accuracy drops to 0.2621 at severity 5 and drastically narrows the recoverable gap. Consequently, a lower recovery ratio for a specific mechanism, such as Laplace+PA, may not stem from a failure of PA to improve utility. Instead, it may result from a disproportionately large gap created by the baseline mechanism's own instability. The recovery ratio metric effectively decouples these distinct sources of variation.

Furthermore, the type of corruption modulates the stability of subspace identification. Brightness corruption yields a nearly constant recovery ratio across severities, suggesting that luminance shifts leave the downstream model's gradient directions largely intact. In contrast, Fog and Defocus Blur show a modest decline in recovery ratio as severity increases. This might be due to the progressive degradation of spatial frequency content, which creates a mismatch between public and private feature distributions. Gaussian Noise, conversely, exhibits an apparent increase in the recovery ratio at higher severities. This is not indicative of enhanced PA effectiveness. Rather, it reflects the rapid contraction of the recoverable gap as no-randomization accuracy collapses, which mechanically inflates the ratio even when absolute gains remain stable.

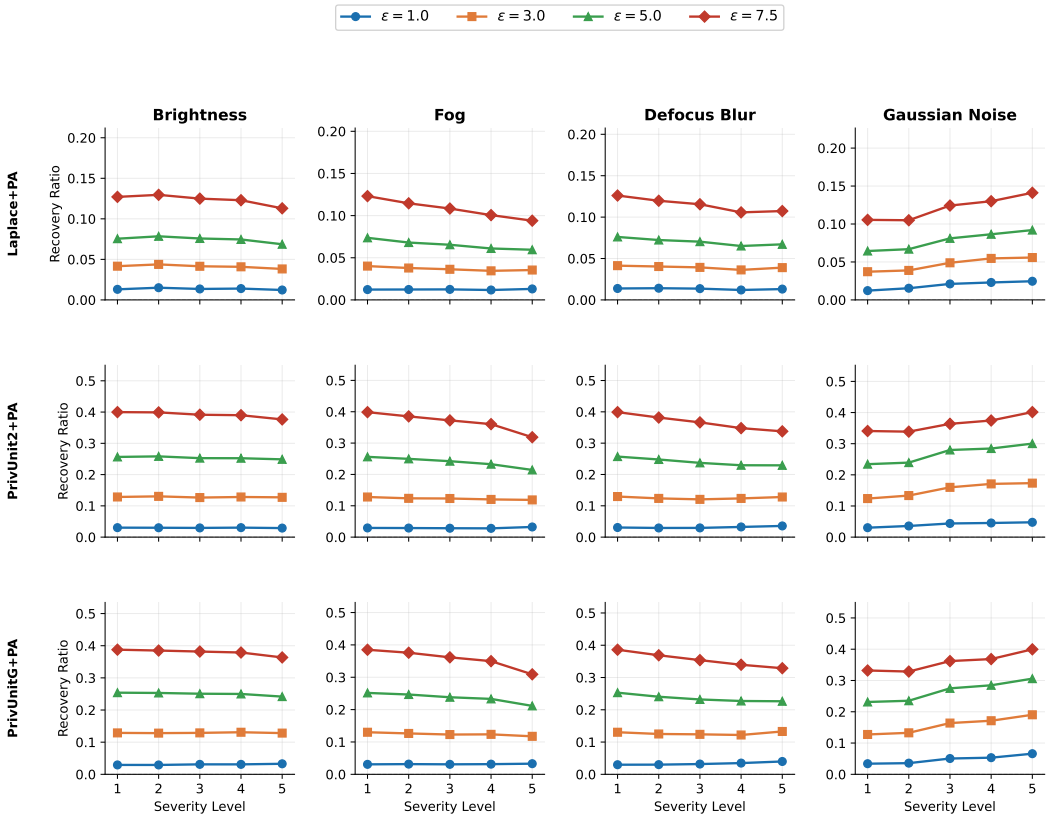


Figure 8: **Recovery Ratio of PA across corruption types, severity levels, and privacy budgets.** Recovery Ratio is defined as $(\text{Acc(Baseline+PA)} - \text{Acc(Baseline)})/(\text{Acc(No Randomization)} - \text{Acc(Baseline)})$. Results are shown for Laplace, PrivUnit2, and PrivUnitG baselines (rows) on four CIFAR-10-C corruption types (columns) at severity levels 1–5, with $\epsilon \in 1.0, 3.0, 5.0, 7.5$. A higher Recovery Ratio indicates that PA restores a greater fraction of the accuracy degraded by LDP noise, independent of the absolute difficulty imposed by the corruption.

Table 6: Main results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10** and **CIFAR-10-C (Brightness)** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

		Privacy Budget (ϵ)											
Mechanism	S	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937
	1	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909	0.8909
	2	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856	0.8856
	3	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783	0.8783
	4	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661	0.8661
	5	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365	0.8365
Laplace (ℓ_1)	0	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1059 (0.0027)	0.1074 (0.0026)	0.1088 (0.0025)	0.1103 (0.0027)	0.1119 (0.0026)	0.1133 (0.0025)	0.1212 (0.0026)	0.1301 (0.0024)
	1	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1060 (0.0027)	0.1073 (0.0026)	0.1088 (0.0026)	0.1103 (0.0027)	0.1119 (0.0026)	0.1134 (0.0025)	0.1212 (0.0026)	0.1301 (0.0025)
	2	0.1002 (0.0029)	0.1016 (0.0029)	0.1030 (0.0029)	0.1045 (0.0028)	0.1059 (0.0027)	0.1073 (0.0026)	0.1086 (0.0027)	0.1102 (0.0026)	0.1116 (0.0026)	0.1131 (0.0026)	0.1209 (0.0026)	0.1295 (0.0026)
	3	0.1002 (0.0029)	0.1016 (0.0029)	0.1030 (0.0029)	0.1044 (0.0028)	0.1058 (0.0028)	0.1071 (0.0026)	0.1084 (0.0026)	0.1099 (0.0026)	0.1113 (0.0026)	0.1128 (0.0026)	0.1203 (0.0027)	0.1287 (0.0026)
	4	0.1001 (0.0029)	0.1015 (0.0029)	0.1028 (0.0028)	0.1042 (0.0028)	0.1056 (0.0028)	0.1068 (0.0027)	0.1081 (0.0026)	0.1096 (0.0026)	0.1109 (0.0026)	0.1124 (0.0025)	0.1196 (0.0026)	0.1276 (0.0024)
	5	0.1001 (0.0029)	0.1013 (0.0028)	0.1025 (0.0028)	0.1038 (0.0028)	0.1050 (0.0028)	0.1063 (0.0027)	0.1074 (0.0026)	0.1087 (0.0026)	0.1099 (0.0026)	0.1111 (0.0025)	0.1177 (0.0027)	0.1251 (0.0025)
Laplace+PA	0	0.1062 (0.0024)	0.1122 (0.0022)	0.1188 (0.0023)	0.1258 (0.0025)	0.1329 (0.0023)	0.1404 (0.0023)	0.1480 (0.0024)	0.1563 (0.0023)	0.1650 (0.0024)	0.1737 (0.0024)	0.2204 (0.0038)	0.2723 (0.0046)
	1	0.1059 (0.0024)	0.1120 (0.0027)	0.1187 (0.0027)	0.1256 (0.0025)	0.1328 (0.0031)	0.1399 (0.0032)	0.1475 (0.0030)	0.1554 (0.0033)	0.1637 (0.0034)	0.1721 (0.0034)	0.2190 (0.0036)	0.2703 (0.0045)
	2	0.1071 (0.0023)	0.1134 (0.0023)	0.1201 (0.0022)	0.1270 (0.0024)	0.1340 (0.0025)	0.1414 (0.0030)	0.1489 (0.0033)	0.1568 (0.0034)	0.1651 (0.0035)	0.1737 (0.0035)	0.2200 (0.0044)	0.2704 (0.0054)
	3	0.1063 (0.0022)	0.1121 (0.0021)	0.1184 (0.0022)	0.1248 (0.0023)	0.1321 (0.0026)	0.1391 (0.0026)	0.1466 (0.0025)	0.1545 (0.0026)	0.1628 (0.0026)	0.1708 (0.0026)	0.2150 (0.0034)	0.2650 (0.0046)
	4	0.1061 (0.0028)	0.1122 (0.0026)	0.1181 (0.0027)	0.1241 (0.0031)	0.1307 (0.0034)	0.1378 (0.0034)	0.1452 (0.0032)	0.1528 (0.0034)	0.1605 (0.0036)	0.1686 (0.0036)	0.2114 (0.0050)	0.2583 (0.0044)
	5	0.1048 (0.0033)	0.1103 (0.0033)	0.1161 (0.0033)	0.1219 (0.0036)	0.1280 (0.0034)	0.1342 (0.0034)	0.1405 (0.0035)	0.1467 (0.0037)	0.1538 (0.0039)	0.1609 (0.0037)	0.1989 (0.0045)	0.2416 (0.0050)
PrivUnit2	0	0.1107 (0.0032)	0.1221 (0.0032)	0.1348 (0.0032)	0.1481 (0.0031)	0.1620 (0.0028)	0.1762 (0.0029)	0.1913 (0.0038)	0.2063 (0.0044)	0.2229 (0.0044)	0.2383 (0.0040)	0.3185 (0.0042)	0.3978 (0.0035)
	1	0.1105 (0.0030)	0.1219 (0.0030)	0.1346 (0.0031)	0.1480 (0.0034)	0.1620 (0.0029)	0.1761 (0.0029)	0.1911 (0.0038)	0.2059 (0.0045)	0.2224 (0.0046)	0.2378 (0.0041)	0.3179 (0.0044)	0.3962 (0.0035)
	2	0.1104 (0.0032)	0.1216 (0.0031)	0.1342 (0.0029)	0.1475 (0.0032)	0.1613 (0.0028)	0.1754 (0.0026)	0.1901 (0.0033)	0.2048 (0.0041)	0.2211 (0.0043)	0.2362 (0.0040)	0.3154 (0.0047)	0.3930 (0.0039)
	3	0.1103 (0.0034)	0.1214 (0.0033)	0.1336 (0.0031)	0.1468 (0.0034)	0.1603 (0.0029)	0.1740 (0.0028)	0.1884 (0.0036)	0.2031 (0.0039)	0.2188 (0.0040)	0.2338 (0.0041)	0.3117 (0.0047)	0.3873 (0.0038)
	4	0.1103 (0.0033)	0.1212 (0.0032)	0.1332 (0.0030)	0.1458 (0.0032)	0.1591 (0.0029)	0.1727 (0.0029)	0.1867 (0.0038)	0.2011 (0.0043)	0.2162 (0.0043)	0.2308 (0.0043)	0.3068 (0.0037)	0.3802 (0.0030)
	5	0.1099 (0.0035)	0.1203 (0.0033)	0.1316 (0.0034)	0.1434 (0.0037)	0.1560 (0.0033)	0.1686 (0.0034)	0.1820 (0.0040)	0.1950 (0.0047)	0.2092 (0.0043)	0.2227 (0.0044)	0.2933 (0.0028)	0.3615 (0.0024)
PrivUnit2+PA	0	0.1203 (0.0022)	0.1440 (0.0030)	0.1710 (0.0027)	0.2012 (0.0028)	0.2341 (0.0031)	0.2673 (0.0040)	0.3027 (0.0039)	0.3377 (0.0041)	0.3733 (0.0048)	0.4076 (0.0052)	0.5477 (0.0055)	0.6451 (0.0057)
	1	0.1216 (0.0032)	0.1452 (0.0030)	0.1722 (0.0037)	0.2017 (0.0034)	0.2339 (0.0036)	0.2679 (0.0039)	0.3029 (0.0034)	0.3368 (0.0039)	0.3724 (0.0039)	0.4053 (0.0043)	0.5470 (0.0033)	0.6432 (0.0052)
	2	0.1206 (0.0027)	0.1445 (0.0023)	0.1720 (0.0028)	0.2018 (0.0030)	0.2345 (0.0038)	0.2680 (0.0044)	0.3029 (0.0045)	0.3365 (0.0050)	0.3708 (0.0054)	0.4039 (0.0051)	0.5429 (0.0050)	0.6386 (0.0045)
	3	0.1204 (0.0026)	0.1437 (0.0028)	0.1702 (0.0031)	0.1993 (0.0031)	0.2304 (0.0033)	0.2631 (0.0030)	0.2964 (0.0035)	0.3302 (0.0039)	0.3644 (0.0032)	0.3966 (0.0035)	0.5335 (0.0045)	0.6277 (0.0038)
	4	0.1205 (0.0027)	0.1438 (0.0029)	0.1699 (0.0029)	0.1987 (0.0026)	0.2299 (0.0021)	0.2617 (0.0026)	0.2949 (0.0024)	0.3280 (0.0028)	0.3603 (0.0028)	0.3911 (0.0038)	0.5249 (0.0050)	0.6165 (0.0055)
	5	0.1194 (0.0020)	0.1410 (0.0026)	0.1660 (0.0034)	0.1936 (0.0031)	0.2223 (0.0034)	0.2536 (0.0043)	0.2840 (0.0049)	0.3152 (0.0054)	0.3453 (0.0062)	0.3754 (0.0059)	0.4978 (0.0049)	0.5840 (0.0045)
PrivUnitG	0	0.1100 (0.0027)	0.1212 (0.0030)	0.1335 (0.0028)	0.1464 (0.0035)	0.1600 (0.0037)	0.1741 (0.0040)	0.1887 (0.0044)	0.2036 (0.0039)	0.2187 (0.0040)	0.2342 (0.0043)	0.3103 (0.0034)	0.3685 (0.0035)
	1	0.1097 (0.0030)	0.1210 (0.0028)	0.1335 (0.0028)	0.1465 (0.0034)	0.1599 (0.0038)	0.1738 (0.0040)	0.1885 (0.0046)	0.2034 (0.0042)	0.2182 (0.0042)	0.2336 (0.0045)	0.3100 (0.0037)	0.3679 (0.0033)
	2	0.1096 (0.0029)	0.1207 (0.0026)	0.1332 (0.0028)	0.1460 (0.0036)	0.1594 (0.0037)	0.1731 (0.0038)	0.1877 (0.0044)	0.2024 (0.0039)	0.2172 (0.0043)	0.2324 (0.0042)	0.3076 (0.0041)	0.3653 (0.0036)
	3	0.1096 (0.0029)	0.1206 (0.0027)	0.1326 (0.0028)	0.1453 (0.0033)	0.1587 (0.0041)	0.1723 (0.0038)	0.1864 (0.0043)	0.2007 (0.0041)	0.2152 (0.0042)	0.2301 (0.0044)	0.3037 (0.0041)	0.3605 (0.0032)
	4	0.1092 (0.0029)	0.1202 (0.0031)	0.1321 (0.0029)	0.1443 (0.0032)	0.1577 (0.0040)	0.1709 (0.0040)	0.1848 (0.0042)	0.1986 (0.0041)	0.2127 (0.0040)	0.2276 (0.0041)	0.2986 (0.0039)	0.3539 (0.0035)
	5	0.1083 (0.0030)	0.1190 (0.0031)	0.1299 (0.0031)	0.1415 (0.0032)	0.1541 (0.0041)	0.1665 (0.0040)	0.1798 (0.0042)	0.1931 (0.0038)	0.2061 (0.0038)	0.2197 (0.0040)	0.2858 (0.0039)	0.3375 (0.0034)
PrivUnitG+PA	0	0.1213 (0.0041)	0.1456 (0.0049)	0.1726 (0.0050)	0.2020 (0.0047)	0.2349 (0.0044)	0.2684 (0.0043)	0.3029 (0.0041)	0.3367 (0.0043)	0.3700 (0.0044)	0.4016 (0.0046)	0.5364 (0.0044)	0.6254 (0.0045)
	1	0.1199 (0.0030)	0.1435 (0.0030)	0.1706 (0.0033)	0.2000 (0.0029)	0.2326 (0.0030)	0.2662 (0.0031)	0.3002 (0.0034)	0.3350 (0.0034)	0.3685 (0.0036)	0.4005 (0.0036)	0.5352 (0.0035)	0.6234 (0.0039)
	2	0.1199 (0.0022)	0.1430 (0.0026)	0.1697 (0.0037)	0.1992 (0.0047)	0.2308 (0.0047)	0.2644 (0.0050)	0.2985 (0.0056)	0.3328 (0.0060)	0.3660 (0.0062)	0.3977 (0.0059)	0.5301 (0.0062)	0.6182 (0.0051)
	3	0.1207 (0.0029)	0.1441 (0.0030)	0.1710 (0.0028)	0.1996 (0.0029)	0.2310 (0.0033)	0.2633 (0.0038)	0.2963 (0.0038)	0.3287 (0.0039)	0.3618 (0.0041)	0.3926 (0.0048)	0.5231 (0.0048)	0.6083 (0.0059)
	4	0.1204 (0.0029)	0.1433 (0.0035)	0.1697 (0.0037)	0.1983 (0.0034)	0.2298 (0.0033)	0.2619 (0.0032)	0.2942 (0.0035)	0.3257 (0.0033)	0.3576 (0.0040)	0.3872 (0.0041)	0.5136 (0.0040)	0.5976 (0.0045)
	5	0.1201 (0.0038)	0.1425 (0.0038)	0.1672 (0.0035)	0.1936 (0.0043)	0.2227 (0.0043)	0.2524 (0.0041)	0.2822 (0.0052)	0.3121 (0.0056)	0.3414 (0.0061)	0.3688 (0.0057)	0.4860 (0.0045)	0.5648 (0.0048)
CW+PA w/o bounding	0	0.1039 (0.0024)	0.1077 (0.0025)	0.1120 (0.0026)	0.1161 (0.0026)	0.1205 (0.0026)	0.1251 (0.0027)	0.1298 (0.0026)	0.1346 (0.0025)	0.1397 (0.0026)	0.1446 (0.0028)	0.1716 (0.0027)	0.2008 (0.0026)
	1	0.1038 (0.0030)	0.1075 (0.0032)	0.1115 (0.0030)	0.1156 (0.0028)	0.1200 (0.0029)	0.1245 (0.0027)	0.1291 (0.0029)	0.1338 (0.0029)	0.1386 (0.0030)	0.1434 (0.0029)	0.1698 (0.0038)	0.1999 (0.0041)
	2	0.1050 (0.0022)	0.1088 (0.0022)	0.1128 (0.0022)	0.1174 (0.0023)	0.1215 (0.0024)	0.1261 (0.0025)	0.1305 (0.0031)	0.1353 (0.0031)	0.1399 (0.0032)	0.1449 (0.0034)	0.1715 (0.0035)	0.2003 (0.0041)
	3	0.1041 (0.0026)	0.1076 (0.0025)	0.1114 (0.0026)	0.1155 (0.0027)	0.1196 (0.0028)	0.1239 (0.0030)	0.1285 (0.0029)	0.1329 (0.0031)	0.1380 (0.0029)	0.1429 (0.0032)	0.1686 (0.0033)	0.1961 (0.0035)
	4	0.1046 (0.0032)	0.1082 (0.0032)	0.1119 (0.0031)	0.1159 (0.0030)	0.1197 (0.0031)	0.1238 (0.0034)	0.1281 (0.0033)	0.1325 (0.0032)	0.1370 (0.0032)	0.1416 (0.0033)	0.1662 (0.0040)	0.1934 (0.0050)
	5	0.1025 (0.0032)	0.1058 (0.0030)	0.1092 (0.0031)	0.1128 (0.0030)	0.1164 (0.0033)	0.1201 (0.0036)	0.1241 (0.0039)	0.1282 (0.0037)	0.1323 (0.0038)	0.1366 (0.0037)	0.1581 (0.0042)	0.1828 (0.0040)

Table 7: Main results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10** and **CIFAR-10-C (Fog)** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

		Privacy Budget (ϵ)											
Mechanism	S	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937
	1	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924	0.8924
	2	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701	0.8701
	3	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460	0.8460
	4	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978	0.7978
	5	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138	0.6138
Laplace (ℓ_1)	0	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1059 (0.0027)	0.1074 (0.0026)	0.1088 (0.0025)	0.1103 (0.0027)	0.1119 (0.0026)	0.1133 (0.0025)	0.1212 (0.0026)	0.1301 (0.0024)
	1	0.1002 (0.0028)	0.1016 (0.0029)	0.1030 (0.0028)	0.1044 (0.0028)	0.1058 (0.0027)	0.1072 (0.0026)	0.1085 (0.0026)	0.1100 (0.0025)	0.1115 (0.0027)	0.1131 (0.0026)	0.1207 (0.0027)	0.1291 (0.0026)
	2	0.1001 (0.0029)	0.1014 (0.0029)	0.1026 (0.0029)	0.1038 (0.0027)	0.1052 (0.0028)	0.1064 (0.0027)	0.1076 (0.0027)	0.1091 (0.0026)	0.1105 (0.0027)	0.1118 (0.0026)	0.1187 (0.0026)	0.1263 (0.0026)
	3	0.1000 (0.0029)	0.1011 (0.0029)	0.1023 (0.0029)	0.1034 (0.0028)	0.1046 (0.0028)	0.1058 (0.0028)	0.1069 (0.0027)	0.1082 (0.0025)	0.1095 (0.0026)	0.1108 (0.0025)	0.1171 (0.0027)	0.1240 (0.0027)
	4	0.0999 (0.0029)	0.1009 (0.0029)	0.1020 (0.0029)	0.1029 (0.0028)	0.1039 (0.0026)	0.1049 (0.0027)	0.1060 (0.0027)	0.1072 (0.0027)	0.1082 (0.0026)	0.1094 (0.0027)	0.1149 (0.0027)	0.1210 (0.0027)
	5	0.0996 (0.0029)	0.1003 (0.0029)	0.1010 (0.0029)	0.1017 (0.0028)	0.1024 (0.0027)	0.1033 (0.0027)	0.1041 (0.0026)	0.1047 (0.0026)	0.1054 (0.0026)	0.1063 (0.0026)	0.1105 (0.0028)	0.1146 (0.0028)
Laplace+PA	0	0.1062 (0.0024)	0.1122 (0.0022)	0.1188 (0.0023)	0.1258 (0.0025)	0.1329 (0.0023)	0.1404 (0.0023)	0.1480 (0.0024)	0.1563 (0.0023)	0.1650 (0.0024)	0.1737 (0.0024)	0.2204 (0.0038)	0.2723 (0.0046)
	1	0.1052 (0.0021)	0.1113 (0.0024)	0.1176 (0.0024)	0.1246 (0.0025)	0.1313 (0.0028)	0.1387 (0.0028)	0.1462 (0.0025)	0.1541 (0.0023)	0.1622 (0.0023)	0.1705 (0.0025)	0.2157 (0.0033)	0.2654 (0.0035)
	2	0.1053 (0.0023)	0.1109 (0.0019)	0.1168 (0.0021)	0.1228 (0.0020)	0.1291 (0.0025)	0.1353 (0.0026)	0.1424 (0.0029)	0.1494 (0.0033)	0.1561 (0.0035)	0.1634 (0.0035)	0.2048 (0.0040)	0.2497 (0.0053)
	3	0.1055 (0.0033)	0.1104 (0.0032)	0.1157 (0.0036)	0.1213 (0.0037)	0.1268 (0.0039)	0.1327 (0.0041)	0.1387 (0.0040)	0.1454 (0.0038)	0.1519 (0.0040)	0.1589 (0.0042)	0.1961 (0.0044)	0.2377 (0.0047)
	4	0.1046 (0.0037)	0.1091 (0.0038)	0.1139 (0.0039)	0.1189 (0.0038)	0.1240 (0.0042)	0.1288 (0.0040)	0.1342 (0.0040)	0.1397 (0.0040)	0.1454 (0.0041)	0.1514 (0.0040)	0.1836 (0.0048)	0.2195 (0.0042)
	5	0.1040 (0.0028)	0.1070 (0.0031)	0.1106 (0.0031)	0.1142 (0.0031)	0.1178 (0.0031)	0.1214 (0.0033)	0.1250 (0.0035)	0.1286 (0.0035)	0.1323 (0.0033)	0.1365 (0.0033)	0.1578 (0.0033)	0.1814 (0.0033)
PrivUnit2	0	0.1107 (0.0032)	0.1221 (0.0032)	0.1348 (0.0032)	0.1481 (0.0031)	0.1620 (0.0028)	0.1762 (0.0029)	0.1913 (0.0038)	0.2063 (0.0044)	0.2229 (0.0044)	0.2383 (0.0040)	0.3185 (0.0042)	0.3978 (0.0035)
	1	0.1102 (0.0031)	0.1216 (0.0031)	0.1341 (0.0033)	0.1474 (0.0035)	0.1610 (0.0031)	0.1751 (0.0029)	0.1898 (0.0035)	0.2047 (0.0042)	0.2212 (0.0042)	0.2364 (0.0039)	0.3155 (0.0042)	0.3937 (0.0036)
	2	0.1099 (0.0034)	0.1209 (0.0031)	0.1328 (0.0031)	0.1455 (0.0033)	0.1586 (0.0029)	0.1719 (0.0033)	0.1859 (0.0036)	0.1999 (0.0039)	0.2154 (0.0045)	0.2293 (0.0040)	0.3046 (0.0038)	0.3786 (0.0030)
	3	0.1094 (0.0033)	0.1198 (0.0033)	0.1313 (0.0034)	0.1432 (0.0031)	0.1553 (0.0031)	0.1675 (0.0036)	0.1806 (0.0042)	0.1936 (0.0044)	0.2082 (0.0046)	0.2209 (0.0040)	0.2921 (0.0036)	0.3610 (0.0025)
	4	0.1091 (0.0036)	0.1187 (0.0035)	0.1291 (0.0036)	0.1399 (0.0034)	0.1508 (0.0034)	0.1618 (0.0037)	0.1736 (0.0042)	0.1853 (0.0045)	0.1983 (0.0050)	0.2106 (0.0046)	0.2741 (0.0044)	0.3366 (0.0034)
	5	0.1067 (0.0035)	0.1139 (0.0036)	0.1217 (0.0037)	0.1296 (0.0037)	0.1375 (0.0035)	0.1455 (0.0037)	0.1540 (0.0044)	0.1625 (0.0047)	0.1713 (0.0044)	0.1799 (0.0039)	0.2241 (0.0047)	0.2671 (0.0035)
PrivUnit2+PA	0	0.1203 (0.0022)	0.1440 (0.0030)	0.1710 (0.0027)	0.2012 (0.0028)	0.2341 (0.0031)	0.2673 (0.0040)	0.3027 (0.0039)	0.3377 (0.0041)	0.3733 (0.0048)	0.4076 (0.0052)	0.5477 (0.0055)	0.6451 (0.0057)
	1	0.1202 (0.0028)	0.1442 (0.0024)	0.1707 (0.0027)	0.2007 (0.0029)	0.2332 (0.0032)	0.2669 (0.0032)	0.3014 (0.0038)	0.3356 (0.0052)	0.3701 (0.0052)	0.4044 (0.0047)	0.5455 (0.0044)	0.6413 (0.0044)
	2	0.1206 (0.0032)	0.1426 (0.0036)	0.1684 (0.0033)	0.1961 (0.0036)	0.2263 (0.0033)	0.2583 (0.0039)	0.2918 (0.0041)	0.3242 (0.0050)	0.3572 (0.0043)	0.3893 (0.0044)	0.5224 (0.0036)	0.6164 (0.0047)
	3	0.1185 (0.0024)	0.1404 (0.0029)	0.1651 (0.0030)	0.1921 (0.0039)	0.2204 (0.0041)	0.2511 (0.0046)	0.2820 (0.0041)	0.3119 (0.0047)	0.3427 (0.0053)	0.3723 (0.0054)	0.4984 (0.0052)	0.5885 (0.0044)
	4	0.1175 (0.0036)	0.1377 (0.0036)	0.1608 (0.0036)	0.1855 (0.0039)	0.2115 (0.0036)	0.2383 (0.0036)	0.2663 (0.0033)	0.2941 (0.0042)	0.3212 (0.0058)	0.3472 (0.0057)	0.4629 (0.0052)	0.5453 (0.0043)
	5	0.1149 (0.0026)	0.1301 (0.0030)	0.1469 (0.0030)	0.1643 (0.0029)	0.1822 (0.0039)	0.2011 (0.0035)	0.2194 (0.0043)	0.2374 (0.0044)	0.2552 (0.0047)	0.2729 (0.0048)	0.3483 (0.0047)	0.4033 (0.0057)
PrivUnitG	0	0.1100 (0.0027)	0.1212 (0.0030)	0.1335 (0.0028)	0.1464 (0.0035)	0.1600 (0.0037)	0.1741 (0.0040)	0.1887 (0.0044)	0.2036 (0.0039)	0.2187 (0.0040)	0.2342 (0.0043)	0.3103 (0.0034)	0.3685 (0.0035)
	1	0.1095 (0.0028)	0.1207 (0.0028)	0.1331 (0.0030)	0.1460 (0.0033)	0.1591 (0.0038)	0.1728 (0.0039)	0.1875 (0.0044)	0.2019 (0.0042)	0.2170 (0.0043)	0.2326 (0.0041)	0.3077 (0.0035)	0.3660 (0.0037)
	2	0.1088 (0.0030)	0.1191 (0.0030)	0.1313 (0.0031)	0.1438 (0.0038)	0.1565 (0.0039)	0.1698 (0.0036)	0.1834 (0.0033)	0.1972 (0.0033)	0.2116 (0.0037)	0.2263 (0.0036)	0.2972 (0.0032)	0.3530 (0.0034)
	3	0.1081 (0.0026)	0.1180 (0.0030)	0.1293 (0.0030)	0.1410 (0.0034)	0.1537 (0.0042)	0.1659 (0.0038)	0.1785 (0.0039)	0.1918 (0.0038)	0.2051 (0.0041)	0.2184 (0.0045)	0.2851 (0.0039)	0.3363 (0.0035)
	4	0.1078 (0.0029)	0.1170 (0.0031)	0.1274 (0.0030)	0.1379 (0.0035)	0.1488 (0.0036)	0.1600 (0.0038)	0.1718 (0.0038)	0.1835 (0.0038)	0.1955 (0.0039)	0.2075 (0.0041)	0.2672 (0.0039)	0.3137 (0.0040)
	5	0.1054 (0.0027)	0.1119 (0.0029)	0.1196 (0.0028)	0.1275 (0.0033)	0.1356 (0.0036)	0.1439 (0.0034)	0.1522 (0.0035)	0.1609 (0.0033)	0.1696 (0.0034)	0.1779 (0.0036)	0.2200 (0.0039)	0.2518 (0.0038)
PrivUnitG+PA	0	0.1213 (0.0041)	0.1456 (0.0049)	0.1726 (0.0050)	0.2020 (0.0047)	0.2349 (0.0044)	0.2684 (0.0043)	0.3029 (0.0041)	0.3367 (0.0043)	0.3700 (0.0044)	0.4016 (0.0046)	0.5364 (0.0044)	0.6254 (0.0045)
	1	0.1210 (0.0026)	0.1445 (0.0030)	0.1718 (0.0030)	0.2011 (0.0036)	0.2330 (0.0039)	0.2664 (0.0041)	0.3004 (0.0048)	0.3340 (0.0053)	0.3671 (0.0050)	0.3986 (0.0051)	0.5328 (0.0053)	0.6207 (0.0053)
	2	0.1205 (0.0021)	0.1428 (0.0023)	0.1690 (0.0028)	0.1971 (0.0035)	0.2269 (0.0036)	0.2583 (0.0033)	0.2909 (0.0036)	0.3234 (0.0039)	0.3549 (0.0041)	0.3851 (0.0040)	0.5125 (0.0043)	0.5968 (0.0053)
	3	0.1186 (0.0026)	0.1405 (0.0025)	0.1650 (0.0026)	0.1917 (0.0029)	0.2201 (0.0033)	0.2495 (0.0032)	0.2796 (0.0044)	0.3093 (0.0050)	0.3392 (0.0055)	0.3679 (0.0056)	0.4879 (0.0055)	0.5678 (0.0050)
	4	0.1180 (0.0036)	0.1384 (0.0036)	0.1613 (0.0042)	0.1859 (0.0043)	0.2121 (0.0041)	0.2388 (0.0045)	0.2663 (0.0051)	0.2932 (0.0056)	0.3198 (0.0057)	0.3451 (0.0054)	0.4528 (0.0043)	0.5262 (0.0037)
	5	0.1136 (0.0035)	0.1284 (0.0034)	0.1448 (0.0033)	0.1621 (0.0038)	0.1807 (0.0044)	0.1991 (0.0038)	0.2178 (0.0038)	0.2353 (0.0041)	0.2535 (0.0041)	0.2701 (0.0041)	0.3417 (0.0049)	0.3914 (0.0045)
CW+PA w/o bounding	0	0.1039 (0.0024)	0.1077 (0.0025)	0.1120 (0.0026)	0.1161 (0.0026)	0.1205 (0.0026)	0.1251 (0.0027)	0.1298 (0.0026)	0.1346 (0.0025)	0.1397 (0.0026)	0.1446 (0.0028)	0.1716 (0.0027)	0.2008 (0.0026)
	1	0.1025 (0.0029)	0.1064 (0.0029)	0.1104 (0.0029)	0.1145 (0.0027)	0.1187 (0.0028)	0.1232 (0.0025)	0.1278 (0.0026)	0.1325 (0.0025)	0.1372 (0.0028)	0.1420 (0.0029)	0.1684 (0.0034)	0.1973 (0.0044)
	2	0.1035 (0.0023)	0.1067 (0.0024)	0.1103 (0.0023)	0.1140 (0.0024)	0.1178 (0.0023)	0.1220 (0.0025)	0.1262 (0.0026)	0.1303 (0.0027)	0.1346 (0.0028)	0.1389 (0.0028)	0.1620 (0.0031)	0.1881 (0.0032)
	3	0.1035 (0.0030)	0.1066 (0.0030)	0.1101 (0.0029)	0.1135 (0.0032)	0.1168 (0.0032)	0.1204 (0.0032)	0.1241 (0.0035)	0.1279 (0.0036)	0.1315 (0.0036)	0.1356 (0.0037)	0.1571 (0.0039)	0.1805 (0.0039)
	4	0.1032 (0.0033)	0.1059 (0.0035)	0.1087 (0.0036)	0.1117 (0.0035)	0.1147 (0.0033)	0.1178 (0.0032)	0.1210 (0.0032)	0.1244 (0.0034)	0.1277 (0.0036)	0.1309 (0.0037)	0.1493 (0.0035)	0.1699 (0.0037)
	5	0.1020 (0.0029)	0.1041 (0.0027)	0.1063 (0.0030)	0.1084 (0.0032)	0.1108 (0.0031)	0.1129 (0.0032)	0.1154 (0.0032)	0.1177 (0.0031)	0.1202 (0.0035)	0.1225 (0.0036)	0.1350 (0.0040)	0.1484 (0.0042)

Table 8: Main results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10** and **CIFAR-10-C (Defocus Blur)** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

		Privacy Budget (ϵ)											
Mechanism	S	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937
	1	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905	0.8905
	2	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729	0.8729
	3	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057	0.8057
	4	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922	0.6922
	5	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898	0.4898
Laplace (ℓ_1)	0	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1059 (0.0027)	0.1074 (0.0026)	0.1088 (0.0025)	0.1103 (0.0027)	0.1119 (0.0026)	0.1133 (0.0025)	0.1212 (0.0026)	0.1301 (0.0024)
	1	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0029)	0.1045 (0.0028)	0.1059 (0.0027)	0.1073 (0.0026)	0.1088 (0.0026)	0.1103 (0.0025)	0.1118 (0.0026)	0.1133 (0.0026)	0.1210 (0.0027)	0.1297 (0.0026)
	2	0.1002 (0.0029)	0.1015 (0.0029)	0.1028 (0.0030)	0.1042 (0.0029)	0.1055 (0.0027)	0.1068 (0.0026)	0.1081 (0.0026)	0.1095 (0.0026)	0.1110 (0.0026)	0.1124 (0.0027)	0.1196 (0.0027)	0.1276 (0.0025)
	3	0.1000 (0.0029)	0.1012 (0.0029)	0.1022 (0.0029)	0.1034 (0.0028)	0.1046 (0.0028)	0.1058 (0.0026)	0.1070 (0.0026)	0.1082 (0.0026)	0.1095 (0.0027)	0.1106 (0.0027)	0.1167 (0.0028)	0.1237 (0.0027)
	4	0.0998 (0.0028)	0.1007 (0.0029)	0.1017 (0.0030)	0.1025 (0.0028)	0.1036 (0.0027)	0.1045 (0.0027)	0.1054 (0.0027)	0.1065 (0.0026)	0.1075 (0.0026)	0.1085 (0.0027)	0.1133 (0.0028)	0.1190 (0.0029)
	5	0.0995 (0.0029)	0.1001 (0.0029)	0.1007 (0.0030)	0.1013 (0.0029)	0.1018 (0.0029)	0.1025 (0.0029)	0.1032 (0.0028)	0.1039 (0.0028)	0.1046 (0.0026)	0.1054 (0.0027)	0.1087 (0.0028)	0.1123 (0.0029)
Laplace+PA	0	0.1062 (0.0024)	0.1122 (0.0022)	0.1188 (0.0023)	0.1258 (0.0025)	0.1329 (0.0023)	0.1404 (0.0023)	0.1480 (0.0024)	0.1563 (0.0023)	0.1650 (0.0024)	0.1737 (0.0024)	0.2204 (0.0038)	0.2723 (0.0046)
	1	0.1064 (0.0029)	0.1126 (0.0030)	0.1190 (0.0029)	0.1256 (0.0029)	0.1324 (0.0032)	0.1397 (0.0031)	0.1476 (0.0032)	0.1558 (0.0034)	0.1639 (0.0036)	0.1724 (0.0035)	0.2180 (0.0038)	0.2692 (0.0033)
	2	0.1065 (0.0031)	0.1124 (0.0036)	0.1185 (0.0037)	0.1246 (0.0038)	0.1312 (0.0041)	0.1377 (0.0044)	0.1445 (0.0046)	0.1517 (0.0046)	0.1594 (0.0047)	0.1674 (0.0047)	0.2098 (0.0052)	0.2583 (0.0052)
	3	0.1059 (0.0030)	0.1108 (0.0031)	0.1160 (0.0030)	0.1215 (0.0031)	0.1273 (0.0030)	0.1333 (0.0029)	0.1397 (0.0029)	0.1460 (0.0033)	0.1527 (0.0034)	0.1595 (0.0032)	0.1963 (0.0046)	0.2358 (0.0054)
	4	0.1036 (0.0040)	0.1078 (0.0041)	0.1121 (0.0045)	0.1166 (0.0047)	0.1211 (0.0051)	0.1258 (0.0050)	0.1306 (0.0048)	0.1358 (0.0050)	0.1412 (0.0048)	0.1465 (0.0049)	0.1745 (0.0048)	0.2052 (0.0043)
	5	0.1025 (0.0033)	0.1052 (0.0031)	0.1083 (0.0033)	0.1113 (0.0033)	0.1144 (0.0032)	0.1176 (0.0033)	0.1210 (0.0033)	0.1244 (0.0034)	0.1277 (0.0033)	0.1312 (0.0036)	0.1496 (0.0037)	0.1690 (0.0037)
PrivUnit2	0	0.1107 (0.0032)	0.1221 (0.0032)	0.1348 (0.0032)	0.1481 (0.0031)	0.1620 (0.0028)	0.1762 (0.0029)	0.1913 (0.0038)	0.2063 (0.0044)	0.2229 (0.0044)	0.2383 (0.0040)	0.3185 (0.0042)	0.3978 (0.0035)
	1	0.1105 (0.0031)	0.1220 (0.0030)	0.1347 (0.0031)	0.1479 (0.0034)	0.1617 (0.0028)	0.1759 (0.0029)	0.1910 (0.0035)	0.2059 (0.0040)	0.2224 (0.0041)	0.2375 (0.0036)	0.3176 (0.0037)	0.3966 (0.0039)
	2	0.1098 (0.0033)	0.1209 (0.0031)	0.1331 (0.0031)	0.1463 (0.0036)	0.1596 (0.0033)	0.1733 (0.0035)	0.1877 (0.0042)	0.2020 (0.0044)	0.2177 (0.0043)	0.2322 (0.0037)	0.3094 (0.0045)	0.3841 (0.0036)
	3	0.1094 (0.0038)	0.1194 (0.0035)	0.1304 (0.0036)	0.1422 (0.0035)	0.1541 (0.0035)	0.1662 (0.0039)	0.1789 (0.0044)	0.1917 (0.0045)	0.2052 (0.0044)	0.2177 (0.0041)	0.2849 (0.0043)	0.3510 (0.0038)
	4	0.1082 (0.0034)	0.1166 (0.0031)	0.1259 (0.0033)	0.1357 (0.0031)	0.1456 (0.0031)	0.1557 (0.0032)	0.1665 (0.0038)	0.1769 (0.0039)	0.1881 (0.0044)	0.1984 (0.0038)	0.2528 (0.0044)	0.3062 (0.0042)
	5	0.1057 (0.0038)	0.1118 (0.0039)	0.1183 (0.0038)	0.1252 (0.0036)	0.1325 (0.0037)	0.1392 (0.0038)	0.1463 (0.0035)	0.1532 (0.0038)	0.1606 (0.0037)	0.1678 (0.0034)	0.2023 (0.0040)	0.2362 (0.0034)
PrivUnit2+PA	0	0.1203 (0.0022)	0.1440 (0.0030)	0.1710 (0.0027)	0.2012 (0.0028)	0.2341 (0.0031)	0.2673 (0.0040)	0.3027 (0.0039)	0.3377 (0.0041)	0.3733 (0.0048)	0.4076 (0.0052)	0.5477 (0.0055)	0.6451 (0.0057)
	1	0.1213 (0.0023)	0.1456 (0.0023)	0.1727 (0.0024)	0.2024 (0.0030)	0.2343 (0.0033)	0.2685 (0.0030)	0.3026 (0.0030)	0.3373 (0.0038)	0.3719 (0.0051)	0.4054 (0.0054)	0.5462 (0.0049)	0.6428 (0.0055)
	2	0.1200 (0.0022)	0.1429 (0.0025)	0.1692 (0.0030)	0.1981 (0.0033)	0.2280 (0.0031)	0.2598 (0.0044)	0.2927 (0.0044)	0.3254 (0.0047)	0.3586 (0.0052)	0.3910 (0.0043)	0.5245 (0.0054)	0.6183 (0.0048)
	3	0.1183 (0.0030)	0.1396 (0.0038)	0.1636 (0.0038)	0.1883 (0.0033)	0.2150 (0.0029)	0.2433 (0.0027)	0.2718 (0.0031)	0.3007 (0.0038)	0.3288 (0.0040)	0.3571 (0.0047)	0.4757 (0.0045)	0.5595 (0.0062)
	4	0.1172 (0.0029)	0.1353 (0.0027)	0.1554 (0.0030)	0.1770 (0.0029)	0.1992 (0.0029)	0.2220 (0.0036)	0.2450 (0.0037)	0.2670 (0.0041)	0.2893 (0.0041)	0.3117 (0.0050)	0.4057 (0.0049)	0.4735 (0.0044)
	5	0.1121 (0.0032)	0.1253 (0.0039)	0.1398 (0.0044)	0.1543 (0.0045)	0.1690 (0.0051)	0.1841 (0.0054)	0.1993 (0.0060)	0.2141 (0.0053)	0.2282 (0.0054)	0.2416 (0.0050)	0.2994 (0.0041)	0.3423 (0.0038)
PrivUnitG	0	0.1100 (0.0027)	0.1212 (0.0030)	0.1335 (0.0028)	0.1464 (0.0035)	0.1600 (0.0037)	0.1741 (0.0040)	0.1887 (0.0044)	0.2036 (0.0039)	0.2187 (0.0040)	0.2342 (0.0043)	0.3103 (0.0034)	0.3685 (0.0035)
	1	0.1100 (0.0027)	0.1209 (0.0028)	0.1334 (0.0030)	0.1467 (0.0034)	0.1597 (0.0037)	0.1736 (0.0037)	0.1886 (0.0044)	0.2033 (0.0043)	0.2183 (0.0044)	0.2336 (0.0042)	0.3101 (0.0041)	0.3687 (0.0032)
	2	0.1092 (0.0029)	0.1201 (0.0028)	0.1323 (0.0027)	0.1449 (0.0031)	0.1577 (0.0040)	0.1714 (0.0037)	0.1855 (0.0039)	0.1995 (0.0041)	0.2144 (0.0042)	0.2286 (0.0043)	0.3018 (0.0042)	0.3578 (0.0034)
	3	0.1083 (0.0025)	0.1181 (0.0029)	0.1287 (0.0029)	0.1400 (0.0033)	0.1517 (0.0038)	0.1641 (0.0034)	0.1764 (0.0037)	0.1891 (0.0041)	0.2024 (0.0042)	0.2156 (0.0044)	0.2790 (0.0041)	0.3284 (0.0040)
	4	0.1067 (0.0025)	0.1152 (0.0030)	0.1247 (0.0029)	0.1343 (0.0032)	0.1440 (0.0034)	0.1541 (0.0033)	0.1648 (0.0037)	0.1754 (0.0040)	0.1861 (0.0042)	0.1967 (0.0038)	0.2489 (0.0040)	0.2887 (0.0044)
	5	0.1048 (0.0033)	0.1105 (0.0034)	0.1168 (0.0033)	0.1236 (0.0036)	0.1304 (0.0037)	0.1374 (0.0037)	0.1445 (0.0040)	0.1517 (0.0038)	0.1586 (0.0038)	0.1660 (0.0038)	0.2003 (0.0036)	0.2253 (0.0042)
PrivUnitG+PA	0	0.1213 (0.0041)	0.1456 (0.0049)	0.1726 (0.0050)	0.2020 (0.0047)	0.2349 (0.0044)	0.2684 (0.0043)	0.3029 (0.0041)	0.3367 (0.0043)	0.3700 (0.0044)	0.4016 (0.0046)	0.5364 (0.0044)	0.6254 (0.0045)
	1	0.1201 (0.0033)	0.1438 (0.0034)	0.1713 (0.0033)	0.2007 (0.0035)	0.2332 (0.0038)	0.2671 (0.0041)	0.3011 (0.0042)	0.3349 (0.0047)	0.3680 (0.0042)	0.3998 (0.0053)	0.5342 (0.0055)	0.6230 (0.0042)
	2	0.1195 (0.0030)	0.1427 (0.0033)	0.1685 (0.0038)	0.1971 (0.0039)	0.2275 (0.0045)	0.2591 (0.0049)	0.2908 (0.0052)	0.3231 (0.0052)	0.3540 (0.0065)	0.3835 (0.0066)	0.5125 (0.0060)	0.5982 (0.0063)
	3	0.1194 (0.0035)	0.1400 (0.0038)	0.1639 (0.0035)	0.1891 (0.0034)	0.2158 (0.0034)	0.2436 (0.0035)	0.2722 (0.0040)	0.2999 (0.0048)	0.3266 (0.0050)	0.3525 (0.0051)	0.4654 (0.0054)	0.5407 (0.0055)
	4	0.1168 (0.0020)	0.1353 (0.0023)	0.1547 (0.0028)	0.1757 (0.0030)	0.1974 (0.0035)	0.2197 (0.0038)	0.2428 (0.0035)	0.2662 (0.0032)	0.2882 (0.0036)	0.3093 (0.0033)	0.3994 (0.0031)	0.4603 (0.0038)
	5	0.1125 (0.0032)	0.1256 (0.0037)	0.1399 (0.0034)	0.1546 (0.0036)	0.1695 (0.0040)	0.1842 (0.0036)	0.1984 (0.0039)	0.2121 (0.0046)	0.2261 (0.0051)	0.2393 (0.0045)	0.2955 (0.0046)	0.3348 (0.0047)
CW+PA w/o bounding	0	0.1039 (0.0024)	0.1077 (0.0025)	0.1120 (0.0026)	0.1161 (0.0026)	0.1205 (0.0026)	0.1251 (0.0027)	0.1298 (0.0026)	0.1346 (0.0025)	0.1397 (0.0026)	0.1446 (0.0028)	0.1716 (0.0027)	0.2008 (0.0026)
	1	0.1035 (0.0030)	0.1076 (0.0027)	0.1116 (0.0028)	0.1161 (0.0026)	0.1204 (0.0027)	0.1250 (0.0031)	0.1297 (0.0033)	0.1346 (0.0034)	0.1392 (0.0032)	0.1442 (0.0032)	0.1711 (0.0041)	0.2011 (0.0040)
	2	0.1043 (0.0031)	0.1079 (0.0029)	0.1114 (0.0032)	0.1154 (0.0034)	0.1196 (0.0035)	0.1235 (0.0034)	0.1274 (0.0033)	0.1318 (0.0035)	0.1364 (0.0038)	0.1410 (0.0038)	0.1657 (0.0039)	0.1926 (0.0041)
	3	0.1041 (0.0027)	0.1074 (0.0029)	0.1104 (0.0030)	0.1137 (0.0031)	0.1172 (0.0031)	0.1208 (0.0031)	0.1244 (0.0030)	0.1281 (0.0032)	0.1321 (0.0030)	0.1361 (0.0028)	0.1574 (0.0029)	0.1804 (0.0039)
	4	0.1022 (0.0035)	0.1049 (0.0035)	0.1079 (0.0036)	0.1108 (0.0038)	0.1135 (0.0038)	0.1163 (0.0040)	0.1192 (0.0041)	0.1220 (0.0042)	0.1251 (0.0043)	0.1283 (0.0042)	0.1453 (0.0043)	0.1634 (0.0043)
	5	0.1017 (0.0030)	0.1035 (0.0032)	0.1053 (0.0033)	0.1072 (0.0036)	0.1091 (0.0036)	0.1113 (0.0036)	0.1131 (0.0035)	0.1150 (0.0035)	0.1171 (0.0036)	0.1193 (0.0035)	0.1300 (0.0031)	0.1416 (0.0034)

Table 9: Main results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10** and **CIFAR-10-C (Gaussian Noise)** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

		Privacy Budget (ϵ)											
Mechanism	S	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937
	1	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046	0.7046
	2	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093	0.5093
	3	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512	0.3512
	4	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040	0.3040
	5	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621	0.2621
Laplace (ℓ_1)	0	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1059 (0.0027)	0.1074 (0.0026)	0.1088 (0.0025)	0.1103 (0.0027)	0.1119 (0.0026)	0.1133 (0.0025)	0.1212 (0.0026)	0.1301 (0.0024)
	1	0.0999 (0.0029)	0.1007 (0.0029)	0.1017 (0.0029)	0.1026 (0.0028)	0.1035 (0.0028)	0.1044 (0.0027)	0.1055 (0.0027)	0.1065 (0.0026)	0.1074 (0.0024)	0.1085 (0.0025)	0.1138 (0.0025)	0.1194 (0.0028)
	2	0.0995 (0.0029)	0.1000 (0.0029)	0.1007 (0.0029)	0.1015 (0.0028)	0.1021 (0.0028)	0.1028 (0.0027)	0.1035 (0.0026)	0.1041 (0.0025)	0.1048 (0.0025)	0.1054 (0.0025)	0.1091 (0.0025)	0.1132 (0.0029)
	3	0.0993 (0.0029)	0.0998 (0.0028)	0.1002 (0.0028)	0.1008 (0.0026)	0.1012 (0.0026)	0.1017 (0.0026)	0.1021 (0.0025)	0.1025 (0.0025)	0.1029 (0.0026)	0.1033 (0.0026)	0.1058 (0.0027)	0.1087 (0.0027)
	4	0.0993 (0.0029)	0.0996 (0.0028)	0.0999 (0.0028)	0.1004 (0.0027)	0.1008 (0.0026)	0.1010 (0.0027)	0.1016 (0.0026)	0.1019 (0.0025)	0.1024 (0.0024)	0.1028 (0.0024)	0.1048 (0.0024)	0.1070 (0.0024)
	5	0.0993 (0.0029)	0.0995 (0.0029)	0.0999 (0.0028)	0.1002 (0.0027)	0.1006 (0.0027)	0.1009 (0.0026)	0.1013 (0.0026)	0.1016 (0.0026)	0.1020 (0.0027)	0.1023 (0.0026)	0.1041 (0.0026)	0.1059 (0.0026)
Laplace+PA	0	0.1062 (0.0024)	0.1122 (0.0022)	0.1188 (0.0023)	0.1258 (0.0025)	0.1329 (0.0023)	0.1404 (0.0023)	0.1480 (0.0024)	0.1563 (0.0023)	0.1650 (0.0024)	0.1737 (0.0024)	0.2204 (0.0038)	0.2723 (0.0046)
	1	0.1037 (0.0023)	0.1081 (0.0024)	0.1126 (0.0027)	0.1171 (0.0028)	0.1219 (0.0029)	0.1267 (0.0032)	0.1316 (0.0029)	0.1365 (0.0029)	0.1416 (0.0030)	0.1469 (0.0030)	0.1761 (0.0036)	0.2074 (0.0040)
	2	0.1034 (0.0031)	0.1063 (0.0033)	0.1093 (0.0035)	0.1123 (0.0033)	0.1154 (0.0037)	0.1186 (0.0037)	0.1220 (0.0040)	0.1252 (0.0041)	0.1287 (0.0039)	0.1324 (0.0037)	0.1511 (0.0044)	0.1718 (0.0045)
	3	0.1029 (0.0027)	0.1051 (0.0026)	0.1072 (0.0026)	0.1094 (0.0028)	0.1116 (0.0032)	0.1139 (0.0031)	0.1163 (0.0028)	0.1186 (0.0026)	0.1211 (0.0027)	0.1234 (0.0028)	0.1363 (0.0028)	0.1490 (0.0030)
	4	0.1024 (0.0020)	0.1043 (0.0022)	0.1063 (0.0021)	0.1081 (0.0025)	0.1104 (0.0025)	0.1121 (0.0028)	0.1140 (0.0026)	0.1160 (0.0025)	0.1181 (0.0026)	0.1202 (0.0026)	0.1307 (0.0031)	0.1415 (0.0039)
	5	0.1022 (0.0033)	0.1035 (0.0035)	0.1048 (0.0035)	0.1065 (0.0034)	0.1083 (0.0034)	0.1099 (0.0032)	0.1117 (0.0032)	0.1136 (0.0031)	0.1153 (0.0034)	0.1170 (0.0034)	0.1264 (0.0033)	0.1357 (0.0035)
PrivUnit2	0	0.1107 (0.0032)	0.1221 (0.0032)	0.1348 (0.0032)	0.1481 (0.0031)	0.1620 (0.0028)	0.1762 (0.0029)	0.1913 (0.0038)	0.2063 (0.0044)	0.2229 (0.0044)	0.2383 (0.0040)	0.3185 (0.0042)	0.3978 (0.0035)
	1	0.1081 (0.0041)	0.1166 (0.0037)	0.1265 (0.0036)	0.1366 (0.0033)	0.1469 (0.0031)	0.1569 (0.0032)	0.1678 (0.0037)	0.1783 (0.0042)	0.1897 (0.0043)	0.2008 (0.0038)	0.2576 (0.0044)	0.3113 (0.0037)
	2	0.1058 (0.0036)	0.1126 (0.0036)	0.1200 (0.0038)	0.1274 (0.0035)	0.1349 (0.0034)	0.1421 (0.0033)	0.1500 (0.0038)	0.1580 (0.0047)	0.1666 (0.0047)	0.1749 (0.0044)	0.2143 (0.0029)	0.2503 (0.0026)
	3	0.1046 (0.0034)	0.1096 (0.0032)	0.1150 (0.0031)	0.1205 (0.0031)	0.1257 (0.0033)	0.1310 (0.0034)	0.1370 (0.0039)	0.1429 (0.0045)	0.1480 (0.0045)	0.1539 (0.0044)	0.1813 (0.0034)	0.2053 (0.0046)
	4	0.1038 (0.0037)	0.1080 (0.0034)	0.1125 (0.0036)	0.1176 (0.0038)	0.1220 (0.0037)	0.1267 (0.0035)	0.1318 (0.0039)	0.1367 (0.0047)	0.1414 (0.0047)	0.1465 (0.0044)	0.1694 (0.0033)	0.1887 (0.0039)
	5	0.1029 (0.0038)	0.1067 (0.0038)	0.1107 (0.0039)	0.1150 (0.0038)	0.1187 (0.0038)	0.1225 (0.0037)	0.1268 (0.0041)	0.1311 (0.0046)	0.1350 (0.0048)	0.1392 (0.0044)	0.1575 (0.0027)	0.1746 (0.0032)
PrivUnit2+PA	0	0.1203 (0.0022)	0.1440 (0.0030)	0.1710 (0.0027)	0.2012 (0.0028)	0.2341 (0.0031)	0.2673 (0.0040)	0.3027 (0.0039)	0.3377 (0.0041)	0.3733 (0.0048)	0.4076 (0.0052)	0.5477 (0.0055)	0.6451 (0.0057)
	1	0.1159 (0.0030)	0.1344 (0.0035)	0.1551 (0.0038)	0.1774 (0.0041)	0.2007 (0.0036)	0.2247 (0.0033)	0.2491 (0.0033)	0.2734 (0.0042)	0.2961 (0.0039)	0.3188 (0.0039)	0.4099 (0.0057)	0.4742 (0.0040)
	2	0.1125 (0.0019)	0.1268 (0.0024)	0.1419 (0.0027)	0.1580 (0.0030)	0.1744 (0.0029)	0.1911 (0.0027)	0.2077 (0.0024)	0.2240 (0.0028)	0.2395 (0.0028)	0.2549 (0.0026)	0.3142 (0.0040)	0.3528 (0.0046)
	3	0.1096 (0.0025)	0.1202 (0.0030)	0.1315 (0.0035)	0.1429 (0.0035)	0.1545 (0.0040)	0.1662 (0.0041)	0.1777 (0.0038)	0.1884 (0.0037)	0.1983 (0.0029)	0.2091 (0.0033)	0.2431 (0.0038)	0.2646 (0.0042)
	4	0.1080 (0.0034)	0.1169 (0.0030)	0.1263 (0.0030)	0.1363 (0.0031)	0.1467 (0.0036)	0.1570 (0.0039)	0.1661 (0.0036)	0.1749 (0.0037)	0.1830 (0.0031)	0.1913 (0.0020)	0.2198 (0.0023)	0.2356 (0.0036)
	5	0.1067 (0.0039)	0.1141 (0.0036)	0.1218 (0.0041)	0.1299 (0.0039)	0.1385 (0.0036)	0.1467 (0.0040)	0.1545 (0.0037)	0.1618 (0.0038)	0.1693 (0.0033)	0.1761 (0.0027)	0.1995 (0.0024)	0.2126 (0.0035)
PrivUnitG	0	0.1100 (0.0027)	0.1212 (0.0030)	0.1335 (0.0028)	0.1464 (0.0035)	0.1600 (0.0037)	0.1741 (0.0040)	0.1887 (0.0044)	0.2036 (0.0039)	0.2187 (0.0040)	0.2342 (0.0043)	0.3103 (0.0034)	0.3685 (0.0035)
	1	0.1071 (0.0028)	0.1158 (0.0029)	0.1250 (0.0034)	0.1348 (0.0039)	0.1448 (0.0040)	0.1553 (0.0039)	0.1658 (0.0040)	0.1764 (0.0043)	0.1873 (0.0047)	0.1983 (0.0043)	0.2513 (0.0035)	0.2920 (0.0037)
	2	0.1051 (0.0028)	0.1118 (0.0031)	0.1190 (0.0032)	0.1264 (0.0036)	0.1336 (0.0040)	0.1415 (0.0041)	0.1493 (0.0043)	0.1569 (0.0043)	0.1646 (0.0043)	0.1726 (0.0046)	0.2101 (0.0046)	0.2378 (0.0045)
	3	0.1031 (0.0026)	0.1078 (0.0023)	0.1132 (0.0025)	0.1188 (0.0028)	0.1242 (0.0029)	0.1297 (0.0034)	0.1352 (0.0039)	0.1405 (0.0046)	0.1458 (0.0045)	0.1511 (0.0046)	0.1772 (0.0051)	0.1965 (0.0044)
	4	0.1024 (0.0028)	0.1067 (0.0029)	0.1113 (0.0030)	0.1162 (0.0032)	0.1209 (0.0034)	0.1254 (0.0036)	0.1298 (0.0041)	0.1345 (0.0041)	0.1390 (0.0041)	0.1438 (0.0044)	0.1656 (0.0035)	0.1807 (0.0035)
	5	0.1021 (0.0026)	0.1059 (0.0027)	0.1100 (0.0028)	0.1139 (0.0032)	0.1175 (0.0035)	0.1213 (0.0037)	0.1250 (0.0037)	0.1290 (0.0035)	0.1332 (0.0039)	0.1372 (0.0043)	0.1550 (0.0040)	0.1681 (0.0037)
PrivUnitG+PA	0	0.1213 (0.0041)	0.1456 (0.0049)	0.1726 (0.0050)	0.2020 (0.0047)	0.2349 (0.0044)	0.2684 (0.0043)	0.3029 (0.0041)	0.3367 (0.0043)	0.3700 (0.0044)	0.4016 (0.0046)	0.5364 (0.0044)	0.6254 (0.0045)
	1	0.1174 (0.0025)	0.1358 (0.0029)	0.1570 (0.0028)	0.1787 (0.0035)	0.2017 (0.0041)	0.2253 (0.0045)	0.2491 (0.0052)	0.2723 (0.0050)	0.2938 (0.0052)	0.3154 (0.0051)	0.4018 (0.0059)	0.4619 (0.0051)
	2	0.1121 (0.0034)	0.1260 (0.0036)	0.1415 (0.0034)	0.1577 (0.0031)	0.1740 (0.0029)	0.1903 (0.0029)	0.2069 (0.0030)	0.2230 (0.0026)	0.2383 (0.0028)	0.2518 (0.0026)	0.3084 (0.0034)	0.3454 (0.0038)
	3	0.1100 (0.0038)	0.1201 (0.0039)	0.1316 (0.0042)	0.1430 (0.0033)	0.1542 (0.0028)	0.1660 (0.0031)	0.1775 (0.0032)	0.1872 (0.0033)	0.1971 (0.0032)	0.2061 (0.0036)	0.2402 (0.0030)	0.2625 (0.0030)
	4	0.1081 (0.0027)	0.1172 (0.0029)	0.1271 (0.0034)	0.1366 (0.0035)	0.1467 (0.0036)	0.1560 (0.0038)	0.1650 (0.0040)	0.1735 (0.0036)	0.1822 (0.0036)	0.1894 (0.0032)	0.2166 (0.0019)	0.2345 (0.0028)
	5	0.1081 (0.0015)	0.1162 (0.0020)	0.1242 (0.0026)	0.1317 (0.0024)	0.1403 (0.0023)	0.1481 (0.0021)	0.1556 (0.0022)	0.1626 (0.0019)	0.1691 (0.0020)	0.1754 (0.0019)	0.1978 (0.0028)	0.2118 (0.0028)
CW+PA w/o bounding	0	0.1039 (0.0024)	0.1077 (0.0025)	0.1120 (0.0026)	0.1161 (0.0026)	0.1205 (0.0026)	0.1251 (0.0027)	0.1298 (0.0026)	0.1346 (0.0025)	0.1397 (0.0026)	0.1446 (0.0028)	0.1716 (0.0027)	0.2008 (0.0026)
	1	0.1027 (0.0016)	0.1051 (0.0025)	0.1079 (0.0027)	0.1106 (0.0029)	0.1133 (0.0029)	0.1161 (0.0029)	0.1191 (0.0028)	0.1221 (0.0030)	0.1252 (0.0029)	0.1286 (0.0028)	0.1452 (0.0028)	0.1636 (0.0030)
	2	0.1022 (0.0034)	0.1041 (0.0034)	0.1059 (0.0036)	0.1077 (0.0036)	0.1098 (0.0035)	0.1119 (0.0037)	0.1138 (0.0037)	0.1159 (0.0038)	0.1180 (0.0036)	0.1201 (0.0037)	0.1309 (0.0038)	0.1429 (0.0039)
	3	0.1019 (0.0023)	0.1034 (0.0023)	0.1048 (0.0023)	0.1060 (0.0024)	0.1075 (0.0027)	0.1089 (0.0027)	0.1103 (0.0028)	0.1118 (0.0028)	0.1133 (0.0028)	0.1150 (0.0028)	0.1229 (0.0031)	0.1313 (0.0032)
	4	0.1021 (0.0027)	0.1033 (0.0028)	0.1043 (0.0026)	0.1055 (0.0026)	0.1065 (0.0026)	0.1077 (0.0027)	0.1089 (0.0026)	0.1102 (0.0024)	0.1114 (0.0023)	0.1126 (0.0024)	0.1191 (0.0030)	0.1263 (0.0032)
	5	0.1017 (0.0029)	0.1027 (0.0029)	0.1037 (0.0028)	0.1047 (0.0029)	0.1058 (0.0030)	0.1069 (0.0032)	0.1080 (0.0032)	0.1090 (0.0032)	0.1099 (0.0034)	0.1112 (0.0034)	0.1168 (0.0036)	0.1225 (0.0038)

A.10.9 Main Results: Accuracy of Linear Regression, Linear Classification, and Nonlinear Classification on CIFAR-10 and MNIST Test Datasets

We provide our main results for linear regression, linear classification, and nonlinear classification under ϵ-LDP mechanisms. In the tables, PA denotes the proposed approach.

Table 10: Main results for $(m = 16)$ **Linear Regression** on **LHSM** $(m = 16)$ under ϵ-LDP mechanisms. We evaluate utility by measuring the RMSE using real values. Lower RMSE indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691	0.0691
Laplace (ℓ_1)	17.6293 (0.3513)	8.8154 (0.1761)	5.8778 (0.1177)	4.4093 (0.0885)	3.5284 (0.0710)	2.9413 (0.0593)	2.5221 (0.0509)	2.2078 (0.0447)	1.9635 (0.0398)	1.7682 (0.0359)	1.1832 (0.0242)	0.8921 (0.0184)
Laplace+PA	1.0520 (0.0245)	0.5387 (0.0122)	0.3729 (0.0086)	0.2935 (0.0075)	0.2482 (0.0074)	0.2197 (0.0077)	0.2006 (0.0080)	0.1872 (0.0084)	0.1774 (0.0088)	0.1700 (0.0091)	0.1511 (0.0101)	0.1439 (0.0105)
PrivUnit2	3.9709 (0.0484)	2.0035 (0.0248)	1.3492 (0.0144)	1.0252 (0.0131)	0.8348 (0.0108)	0.7097 (0.0087)	0.6222 (0.0093)	0.5577 (0.0077)	0.5083 (0.0082)	0.4710 (0.0064)	0.3619 (0.0062)	0.3155 (0.0058)
PrivUnit2+PA	0.9048 (0.0155)	0.4776 (0.0081)	0.3437 (0.0064)	0.2822 (0.0057)	0.2490 (0.0053)	0.2282 (0.0062)	0.2150 (0.0066)	0.2058 (0.0068)	0.1991 (0.0072)	0.1945 (0.0072)	0.1826 (0.0080)	0.1780 (0.0083)
PrivUnitG	3.9645 (0.0625)	1.9924 (0.0252)	1.3471 (0.0171)	1.0290 (0.0136)	0.8425 (0.0100)	0.7192 (0.0092)	0.6333 (0.0077)	0.5707 (0.0075)	0.5235 (0.0070)	0.4865 (0.0061)	0.3831 (0.0059)	0.3365 (0.0053)
PrivUnitG+PA	0.8996 (0.0106)	0.4736 (0.0069)	0.3406 (0.0055)	0.2801 (0.0048)	0.2471 (0.0051)	0.2269 (0.0054)	0.2136 (0.0061)	0.2044 (0.0065)	0.1981 (0.0072)	0.1936 (0.0071)	0.1822 (0.0081)	0.1765 (0.0084)
CW+PA w/o bounding	1.1897 (0.0281)	0.5976 (0.0142)	0.4016 (0.0096)	0.3045 (0.0073)	0.2470 (0.0060)	0.2093 (0.0051)	0.1828 (0.0045)	0.1634 (0.0040)	0.1486 (0.0037)	0.1371 (0.0035)	0.1048 (0.0031)	0.0908 (0.0030)
Task-Aware	0.1920 (0.0093)	0.1920 (0.0093)	0.1918 (0.0093)	0.1916 (0.0094)	0.1913 (0.0094)	0.1910 (0.0094)	0.1906 (0.0094)	0.1901 (0.0094)	0.1896 (0.0093)	0.1890 (0.0093)	0.1852 (0.0093)	0.1802 (0.0092)

Table 11: Main results for $(m = 64)$ **Linear Classification** on **CIFAR-10** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058	0.9058
Laplace (ℓ_1)	0.1020 (0.0027)	0.1041 (0.0028)	0.1063 (0.0030)	0.1083 (0.0028)	0.1103 (0.0028)	0.1126 (0.0027)	0.1151 (0.0028)	0.1176 (0.0030)	0.1199 (0.0029)	0.1221 (0.0029)	0.1347 (0.0027)	0.1480 (0.0029)
Laplace+PA	0.1085 (0.0033)	0.1170 (0.0035)	0.1258 (0.0037)	0.1352 (0.0037)	0.1454 (0.0035)	0.1555 (0.0036)	0.1661 (0.0038)	0.1777 (0.0038)	0.1891 (0.0041)	0.2006 (0.0035)	0.2658 (0.0038)	0.3342 (0.0045)
PrivUnit2	0.1174 (0.0034)	0.1363 (0.0038)	0.1572 (0.0047)	0.1805 (0.0048)	0.2047 (0.0052)	0.2299 (0.0051)	0.2561 (0.0054)	0.2827 (0.0056)	0.3109 (0.0047)	0.3397 (0.0047)	0.4681 (0.0047)	0.5728 (0.0040)
PrivUnit2+PA	0.1258 (0.0026)	0.1587 (0.0033)	0.1989 (0.0038)	0.2461 (0.0039)	0.2967 (0.0033)	0.3496 (0.0037)	0.4014 (0.0049)	0.4510 (0.0049)	0.4962 (0.0055)	0.5368 (0.0045)	0.6782 (0.0040)	0.7519 (0.0036)
PrivUnitG	0.1161 (0.0037)	0.1353 (0.0041)	0.1564 (0.0044)	0.1795 (0.0047)	0.2032 (0.0050)	0.2279 (0.0055)	0.2546 (0.0057)	0.2808 (0.0054)	0.3067 (0.0059)	0.3330 (0.0059)	0.4551 (0.0054)	0.5392 (0.0053)
PrivUnitG+PA	0.1265 (0.0039)	0.1598 (0.0040)	0.1997 (0.0048)	0.2464 (0.0040)	0.2971 (0.0047)	0.3492 (0.0042)	0.3997 (0.0041)	0.4472 (0.0038)	0.4912 (0.0037)	0.5302 (0.0037)	0.6675 (0.0037)	0.7485 (0.0031)
CW+PA w/o bounding	0.1057 (0.0030)	0.1114 (0.0032)	0.1170 (0.0032)	0.1227 (0.0034)	0.1286 (0.0035)	0.1349 (0.0033)	0.1414 (0.0033)	0.1480 (0.0033)	0.1546 (0.0030)	0.1615 (0.0034)	0.2001 (0.0038)	0.2426 (0.0039)
Task-Aware	0.1000 (0.0000)	0.1000 (0.0000)	0.1001 (0.0002)	0.1002 (0.0004)	0.1005 (0.0005)	0.1011 (0.0007)	0.1018 (0.0008)	0.1028 (0.0011)	0.1039 (0.0013)	0.1053 (0.0017)	0.1145 (0.0014)	0.1276 (0.0022)

Table 12: Main results for **Nonlinear MLP Classifier with ReLU** $(m = 64)$ on **CIFAR-10** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937	0.8937
Laplace (ℓ_1)	0.1002 (0.0029)	0.1017 (0.0029)	0.1031 (0.0028)	0.1046 (0.0028)	0.1059 (0.0027)	0.1074 (0.0026)	0.1088 (0.0025)	0.1103 (0.0027)	0.1119 (0.0026)	0.1133 (0.0025)	0.1212 (0.0026)	0.1301 (0.0024)
Laplace+PA	0.1062 (0.0024)	0.1122 (0.0022)	0.1188 (0.0023)	0.1258 (0.0025)	0.1329 (0.0023)	0.1404 (0.0023)	0.1480 (0.0024)	0.1563 (0.0023)	0.1650 (0.0024)	0.1737 (0.0024)	0.2204 (0.0038)	0.2723 (0.0046)
PrivUnit2	0.1107 (0.0032)	0.1221 (0.0032)	0.1348 (0.0032)	0.1481 (0.0031)	0.1620 (0.0028)	0.1762 (0.0029)	0.1913 (0.0038)	0.2063 (0.0044)	0.2229 (0.0044)	0.2383 (0.0040)	0.3185 (0.0042)	0.3978 (0.0035)
PrivUnit2+PA	0.1203 (0.0022)	0.1440 (0.0030)	0.1710 (0.0027)	0.2012 (0.0028)	0.2341 (0.0031)	0.2673 (0.0040)	0.3027 (0.0039)	0.3377 (0.0041)	0.3733 (0.0048)	0.4076 (0.0052)	0.5477 (0.0055)	0.6451 (0.0057)
PrivUnitG	0.1100 (0.0027)	0.1212 (0.0030)	0.1335 (0.0028)	0.1464 (0.0035)	0.1600 (0.0037)	0.1741 (0.0040)	0.1887 (0.0044)	0.2036 (0.0039)	0.2187 (0.0040)	0.2342 (0.0043)	0.3103 (0.0034)	0.3685 (0.0035)
PrivUnitG+PA	0.1213 (0.0041)	0.1456 (0.0049)	0.1726 (0.0050)	0.2020 (0.0047)	0.2349 (0.0044)	0.2684 (0.0043)	0.3029 (0.0041)	0.3367 (0.0043)	0.3700 (0.0044)	0.4016 (0.0046)	0.5364 (0.0044)	0.6254 (0.0045)
CW+PA w/o bounding	0.1039 (0.0024)	0.1077 (0.0025)	0.1120 (0.0026)	0.1161 (0.0026)	0.1205 (0.0026)	0.1251 (0.0027)	0.1298 (0.0026)	0.1346 (0.0025)	0.1397 (0.0026)	0.1446 (0.0028)	0.1716 (0.0027)	0.2008 (0.0026)

Table 13: Main results for $(m = 64)$ **Nonlinear MLP Classifier with GELU** on **CIFAR-10** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012	0.9012
Laplace (ℓ_1)	0.1015 (0.0033)	0.1021 (0.0034)	0.1027 (0.0033)	0.1033 (0.0032)	0.1039 (0.0033)	0.1044 (0.0033)	0.1050 (0.0032)	0.1055 (0.0032)	0.1061 (0.0031)	0.1068 (0.0032)	0.1100 (0.0033)	0.1134 (0.0036)
Laplace+PA	0.1064 (0.0032)	0.1114 (0.0034)	0.1166 (0.0032)	0.1221 (0.0033)	0.1275 (0.0036)	0.1336 (0.0035)	0.1400 (0.0038)	0.1464 (0.0036)	0.1532 (0.0042)	0.1598 (0.0041)	0.1968 (0.0041)	0.2380 (0.0042)
PrivUnit2	0.1055 (0.0028)	0.1094 (0.0029)	0.1144 (0.0031)	0.1191 (0.0032)	0.1237 (0.0033)	0.1286 (0.0034)	0.1336 (0.0036)	0.1384 (0.0041)	0.1436 (0.0044)	0.1492 (0.0052)	0.1723 (0.0038)	0.1987 (0.0034)
PrivUnit2+PA	0.1172 (0.0032)	0.1375 (0.0027)	0.1606 (0.0029)	0.1852 (0.0030)	0.2115 (0.0032)	0.2400 (0.0035)	0.2692 (0.0034)	0.2985 (0.0037)	0.3276 (0.0048)	0.3565 (0.0054)	0.4834 (0.0052)	0.5766 (0.0053)
PrivUnitG	0.1044 (0.0029)	0.1087 (0.0030)	0.1133 (0.0031)	0.1181 (0.0034)	0.1229 (0.0030)	0.1278 (0.0031)	0.1325 (0.0030)	0.1376 (0.0033)	0.1424 (0.0033)	0.1471 (0.0035)	0.1709 (0.0036)	0.1900 (0.0039)
PrivUnitG+PA	0.1181 (0.0027)	0.1384 (0.0027)	0.1612 (0.0029)	0.1854 (0.0033)	0.2122 (0.0036)	0.2399 (0.0040)	0.2682 (0.0038)	0.2968 (0.0038)	0.3255 (0.0043)	0.3523 (0.0052)	0.4730 (0.0044)	0.5591 (0.0051)
CW+PA w/o bounding	0.1045 (0.0032)	0.1076 (0.0031)	0.1109 (0.0032)	0.1142 (0.0031)	0.1176 (0.0033)	0.1211 (0.0033)	0.1248 (0.0033)	0.1283 (0.0035)	0.1321 (0.0033)	0.1357 (0.0035)	0.1567 (0.0039)	0.1801 (0.0046)

Table 14: Main results for $(m = 64)$ **Nonlinear MLP Classifier with Tanh** on **CIFAR-10** under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy. Higher accuracy indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014	0.9014
Laplace (ℓ_1)	0.1013 (0.0022)	0.1025 (0.0022)	0.1035 (0.0023)	0.1046 (0.0022)	0.1057 (0.0022)	0.1069 (0.0023)	0.1079 (0.0022)	0.1091 (0.0022)	0.1102 (0.0021)	0.1113 (0.0021)	0.1171 (0.0021)	0.1237 (0.0025)
Laplace+PA	0.1064 (0.0030)	0.1133 (0.0031)	0.1201 (0.0033)	0.1278 (0.0033)	0.1353 (0.0035)	0.1437 (0.0038)	0.1524 (0.0038)	0.1615 (0.0045)	0.1714 (0.0045)	0.1816 (0.0047)	0.2374 (0.0044)	0.3008 (0.0047)
PrivUnit2	0.1085 (0.0031)	0.1176 (0.0030)	0.1270 (0.0025)	0.1367 (0.0028)	0.1472 (0.0028)	0.1576 (0.0032)	0.1682 (0.0033)	0.1789 (0.0027)	0.1909 (0.0028)	0.2026 (0.0032)	0.2607 (0.0048)	0.3219 (0.0056)
PrivUnit2+PA	0.1215 (0.0032)	0.1491 (0.0038)	0.1829 (0.0038)	0.2217 (0.0045)	0.2646 (0.0048)	0.3106 (0.0036)	0.3587 (0.0034)	0.4059 (0.0040)	0.4532 (0.0051)	0.4973 (0.0051)	0.6615 (0.0042)	0.7488 (0.0037)
PrivUnitG	0.1086 (0.0027)	0.1172 (0.0031)	0.1264 (0.0034)	0.1362 (0.0034)	0.1462 (0.0035)	0.1569 (0.0032)	0.1671 (0.0034)	0.1780 (0.0034)	0.1887 (0.0035)	0.1997 (0.0040)	0.2537 (0.0043)	0.2985 (0.0037)
PrivUnitG+PA	0.1240 (0.0027)	0.1510 (0.0033)	0.1837 (0.0035)	0.2216 (0.0051)	0.2637 (0.0051)	0.3090 (0.0054)	0.3562 (0.0057)	0.4025 (0.0054)	0.4476 (0.0051)	0.4894 (0.0047)	0.6495 (0.0049)	0.7423 (0.0041)
CW+PA w/o bounding	0.1039 (0.0029)	0.1080 (0.0029)	0.1124 (0.0029)	0.1168 (0.0031)	0.1217 (0.0032)	0.1263 (0.0032)	0.1316 (0.0032)	0.1367 (0.0032)	0.1419 (0.0032)	0.1473 (0.0033)	0.1773 (0.0039)	0.2119 (0.0044)

Table 15: Main results for **Nonlinear MLP Classifier with ReLU** on **MNIST (VAE)** $(m = 32)$ under ϵ-LDP mechanisms. We evaluate utility by measuring the classification accuracy using ground-truth labels. Higher accuracy indicates higher utility.

	Privacy Budget (ϵ)											
Mechanism	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
No randomization	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)	0.9829 (0.0000)
Laplace (ℓ_1)	0.1050 (0.0032)	0.1078 (0.0034)	0.1108 (0.0034)	0.1139 (0.0035)	0.1171 (0.0034)	0.1202 (0.0035)	0.1234 (0.0037)	0.1267 (0.0039)	0.1302 (0.0040)	0.1336 (0.0039)	0.1524 (0.0039)	0.1732 (0.0041)
Laplace+PA	0.1092 (0.0030)	0.1178 (0.0032)	0.1271 (0.0029)	0.1369 (0.0029)	0.1474 (0.0032)	0.1584 (0.0035)	0.1704 (0.0037)	0.1828 (0.0040)	0.1956 (0.0039)	0.2091 (0.0037)	0.2846 (0.0037)	0.3666 (0.0041)
PrivUnit2	0.1152 (0.0044)	0.1307 (0.0043)	0.1474 (0.0041)	0.1659 (0.0047)	0.1855 (0.0044)	0.2063 (0.0046)	0.2289 (0.0046)	0.2526 (0.0051)	0.2767 (0.0054)	0.3021 (0.0054)	0.4360 (0.0047)	0.5729 (0.0027)
PrivUnit2+PA	0.1250 (0.0025)	0.1536 (0.0032)	0.1875 (0.0038)	0.2263 (0.0038)	0.2696 (0.0040)	0.3174 (0.0036)	0.3674 (0.0039)	0.4203 (0.0038)	0.4720 (0.0041)	0.5214 (0.0043)	0.7252 (0.0036)	0.8427 (0.0032)
PrivUnitG	0.1139 (0.0034)	0.1290 (0.0034)	0.1462 (0.0037)	0.1640 (0.0032)	0.1822 (0.0029)	0.2014 (0.0028)	0.2217 (0.0031)	0.2424 (0.0035)	0.2632 (0.0038)	0.2838 (0.0043)	0.3862 (0.0054)	0.4682 (0.0053)
PrivUnitG+PA	0.1242 (0.0027)	0.1528 (0.0025)	0.1862 (0.0035)	0.2236 (0.0040)	0.2646 (0.0044)	0.3083 (0.0047)	0.3538 (0.0051)	0.3989 (0.0059)	0.4432 (0.0066)	0.4861 (0.0057)	0.6578 (0.0042)	0.7647 (0.0037)
CW+PA w/o bounding	0.1069 (0.0027)	0.1127 (0.0027)	0.1191 (0.0025)	0.1253 (0.0027)	0.1323 (0.0028)	0.1392 (0.0029)	0.1466 (0.0031)	0.1542 (0.0032)	0.1623 (0.0032)	0.1706 (0.0033)	0.2173 (0.0038)	0.2685 (0.0043)

A.10.10 SUPPLEMENTAL RESULTS: SENSITIVITY ANALYSIS OF THE FINITE SCALE CAP

This analysis is intended as a robustness check rather than a tuning procedure. Since $\lambda_{\max}$ is a public configuration, it does not affect the privacy parameters, but it must be fixed from public information in practice.

We provide our sensitivity analysis on $\lambda_{\max}$ using CIFAR-10-C (Brightness, Severity Level 5). In the tables, PA denotes the proposed approach.

To provide further clarity, we conducted an additional sensitivity analysis of the finite cap $\lambda_{\max}$ used for near-null directions. For Laplace+PA at $\epsilon = 10$, the only visibly lower mean occurs at $\lambda_{\max} = 10$. For $\lambda_{\max} \geq 10^2$, accuracy ranges from 0.2376 to 0.2433, a difference of 0.0057, while the standard deviations across 20 seeds range from approximately 0.0035 to 0.0047. PrivUnit2+PA and PrivUnitG+PA likewise exhibit little variation across the tested range.

Table 16: Sensitivity analysis on the $\lambda_{\max}$ for **Laplace+PA**, **PrivUnit2+PA**, and **PrivUnitG+PA** on **CIFAR-10-C (Brightness Corruption At The Highest Severity Level 5)** Test Dataset under ϵ-LDP mechanisms. We evaluate utility by measuring the Accuracy.

		Privacy Budget (ϵ)											
Mechanism	$\lambda_{\max}$	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace+PA	10^1	0.1044 (0.0038)	0.1091 (0.0039)	0.1141 (0.0039)	0.1195 (0.0036)	0.1251 (0.0035)	0.1308 (0.0035)	0.1365 (0.0035)	0.1424 (0.0036)	0.1485 (0.0037)	0.1552 (0.0040)	0.1891 (0.0038)	0.2264 (0.0045)
	10^2	0.1045 (0.0022)	0.1095 (0.0021)	0.1152 (0.0022)	0.1208 (0.0018)	0.1267 (0.0019)	0.1327 (0.0024)	0.1390 (0.0025)	0.1454 (0.0027)	0.1517 (0.0027)	0.1589 (0.0027)	0.1963 (0.0036)	0.2376 (0.0035)
	10^3	0.1052 (0.0031)	0.1104 (0.0031)	0.1160 (0.0031)	0.1220 (0.0032)	0.1278 (0.0034)	0.1343 (0.0036)	0.1406 (0.0038)	0.1470 (0.0040)	0.1538 (0.0041)	0.1609 (0.0043)	0.1987 (0.0045)	0.2401 (0.0047)
	10^4	0.1049 (0.0032)	0.1101 (0.0035)	0.1157 (0.0034)	0.1216 (0.0036)	0.1277 (0.0036)	0.1341 (0.0034)	0.1404 (0.0033)	0.1473 (0.0033)	0.1541 (0.0034)	0.1612 (0.0036)	0.2002 (0.0034)	0.2424 (0.0037)
	10^5	0.1056 (0.0016)	0.1106 (0.0018)	0.1163 (0.0021)	0.1222 (0.0022)	0.1283 (0.0020)	0.1342 (0.0022)	0.1405 (0.0028)	0.1476 (0.0031)	0.1543 (0.0028)	0.1608 (0.0031)	0.1990 (0.0035)	0.2409 (0.0037)
PrivUnit2+PA	10^1	0.1189 (0.0036)	0.1404 (0.0035)	0.1653 (0.0038)	0.1922 (0.0036)	0.2218 (0.0034)	0.2522 (0.0039)	0.2832 (0.0040)	0.3146 (0.0046)	0.3438 (0.0047)	0.3733 (0.0042)	0.4983 (0.0051)	0.5851 (0.0053)
	10^2	0.1187 (0.0031)	0.1408 (0.0034)	0.1663 (0.0037)	0.1940 (0.0038)	0.2236 (0.0036)	0.2537 (0.0040)	0.2846 (0.0043)	0.3160 (0.0047)	0.3459 (0.0047)	0.3745 (0.0051)	0.4989 (0.0051)	0.5855 (0.0053)
	10^3	0.1187 (0.0025)	0.1409 (0.0028)	0.1660 (0.0034)	0.1943 (0.0032)	0.2241 (0.0036)	0.2549 (0.0033)	0.2865 (0.0034)	0.3177 (0.0028)	0.3483 (0.0037)	0.3763 (0.0038)	0.5012 (0.0043)	0.5851 (0.0062)
	10^4	0.1205 (0.0032)	0.1423 (0.0029)	0.1680 (0.0027)	0.1953 (0.0030)	0.2246 (0.0037)	0.2548 (0.0036)	0.2857 (0.0038)	0.3161 (0.0030)	0.3457 (0.0027)	0.3745 (0.0029)	0.4986 (0.0057)	0.5852 (0.0043)
	10^5	0.1189 (0.0041)	0.1416 (0.0038)	0.1671 (0.0037)	0.1946 (0.0038)	0.2242 (0.0044)	0.2541 (0.0049)	0.2852 (0.0053)	0.3159 (0.0054)	0.3464 (0.0054)	0.3755 (0.0055)	0.4975 (0.0047)	0.5845 (0.0045)
PrivUnitG+PA	10^1	0.1195 (0.0033)	0.1415 (0.0031)	0.1665 (0.0035)	0.1940 (0.0033)	0.2234 (0.0035)	0.2538 (0.0042)	0.2842 (0.0036)	0.3142 (0.0035)	0.3430 (0.0033)	0.3716 (0.0032)	0.4889 (0.0050)	0.5658 (0.0052)
	10^2	0.1198 (0.0031)	0.1414 (0.0032)	0.1665 (0.0038)	0.1931 (0.0034)	0.2215 (0.0038)	0.2514 (0.0040)	0.2814 (0.0037)	0.3117 (0.0038)	0.3410 (0.0039)	0.3688 (0.0032)	0.4874 (0.0046)	0.5652 (0.0041)
	10^3	0.1192 (0.0031)	0.1411 (0.0032)	0.1666 (0.0032)	0.1935 (0.0030)	0.2221 (0.0027)	0.2522 (0.0030)	0.2822 (0.0038)	0.3118 (0.0044)	0.3410 (0.0045)	0.3690 (0.0045)	0.4882 (0.0040)	0.5667 (0.0043)
	10^4	0.1205 (0.0024)	0.1426 (0.0029)	0.1673 (0.0025)	0.1942 (0.0029)	0.2226 (0.0030)	0.2525 (0.0031)	0.2832 (0.0041)	0.3122 (0.0044)	0.3415 (0.0043)	0.3692 (0.0048)	0.4877 (0.0063)	0.5656 (0.0058)
	10^5	0.1198 (0.0022)	0.1414 (0.0026)	0.1662 (0.0032)	0.1932 (0.0027)	0.2216 (0.0027)	0.2514 (0.0035)	0.2819 (0.0039)	0.3115 (0.0042)	0.3413 (0.0042)	0.3692 (0.0043)	0.4866 (0.0032)	0.5644 (0.0036)

A.10.11 Supplemental Results from Agreement Experiments

We present additional results for nonlinear MLPs under ϵ-LDP mechanisms with server-side aggregation. While LDP offers strong privacy guarantees, applying these mechanisms directly to individual representation vectors introduces substantial noise, resulting in insufficient performance on downstream tasks. To improve utility at scale, we therefore adopt a server-side aggregation protocol in which clients independently randomize their local representation vectors $\boldsymbol{z}$ under LDP, and the server computes the empirical mean of the randomized reports. Under independent randomization with bounded noise variance, averaging K reports reduces the variance of the aggregated noise at a rate of $\mathcal{O}(K^{-1})$, improving the performance of the aggregate for downstream tasks.

Evaluation Protocol. Our experiments focus on image classification, with the aggregate representations serving as inputs to a downstream classifier. Specifically, we consider a scenario depicted in Figure 5 where a central server aggregates protected private representations from K virtual clients. These representations are extracted by processing private test images through a public feature extractor and are protected through a LDP mechanism. For an N-round evaluation, the 10,000 samples in the private test dataset are partitioned into N disjoint subsets and randomly distributed among the clients, regardless of their class labels. For instance, with $N = 20$ and $K = 500$, each client is allocated 20 unique test samples. To avoid privacy budget accumulation, we ensure that each representation vector is transmitted only once across the N rounds. Furthermore, inspired by the *label-anchored* setting (Qin and Bai, 2026), each client privatizes its representation using ϵ_{feat}-LDP and its corresponding class label via Randomized Response (RR) under ϵ_{label}-LDP. The server groups these reports by their *perturbed* labels, computes the mean randomized representation for each class, and subsequently feeds these aggregates into a downstream classifier for prediction.

Utility. We evaluate the utility of our methodologies by measuring the prediction preservation rate (also referred to as the agreement rate); agreement rate and fidelity are commonly used in Knowledge Distillation, where a student model is evaluated against a teacher model's predictions. The prediction preservation rate is calculated as the fraction of class groups for which the classifier's prediction on the noisy aggregate aligns with that on the noiseless aggregate. This formulation isolates the performance degradation attributable solely to the privacy-preserving noise, thereby decoupling it from the inherent performance limitations of the underlying model. A higher preservation rate indicates greater utility, implying that the injected privacy noise does not significantly distort the model's original decision boundaries. Without any randomization, the prediction preservation rate is 1.0 by definition.

Privacy Composition. From a privacy perspective, because each representation is transmitted only once, the budget does not accumulate over multiple rounds. However, under the label-anchored setting, each transmission incurs a total privacy loss of $\epsilon_{\text{feat}} + \epsilon_{\text{label}}$ per item, in accordance with the basic composition theorem.

Results. We examine the effects of label privacy, aggregation size, and mechanism. Tables 18–23 report the averaged prediction preservation rates on CIFAR-10 (ResNet-20) and MNIST (VAE) for $\epsilon_{\text{label}} \in \{1, 2, 3\}$. Our findings are as follows:

First, label privacy strongly affects the benefits of aggregation. With $\epsilon_{\text{label}} = 1$ and $K = 500$, preservation rates remain limited even at $\epsilon_{\text{feat}} = 10$, reaching at most 0.5243 on CIFAR-10 and 0.6620 on MNIST among the evaluated mechanisms. Relaxing label privacy substantially improves preservation. For example, for PrivUnit2+PA at $K = 500$ and $\epsilon_{\text{feat}} = 10$ on CIFAR-10, increasing ϵ_{label} from 1 to 2 and then to 3 raises the preservation rate from 0.5243 to 0.8435 and 0.9995, respectively. A possible explanation is that label perturbation mixes representations from different classes within each perturbed-label group, weakening the distinctions between group means. Increasing the number of reports reduces the variance of independent feature noise but does not, by itself, eliminate this class mixing. These observations suggest that the benefits of aggregation depend on both the reliability of the perturbed labels and the number of reports.

Second, larger aggregates enable high prediction preservation under suitable privacy budgets. The benefits of increasing K are particularly clear at $\epsilon_{\text{label}} = 3$. For PrivUnit2+PA with $\epsilon_{\text{feat}} = 2$, increasing K from 100 to 500 raises the preservation rate from 0.3415 to 0.7548 on CIFAR-10 and from 0.4120 to 0.8590 on MNIST. At $K = 500$ and $\epsilon_{\text{label}} = 3$, the same mechanism achieves preservation rates of at least 0.99 at all evaluated feature budgets of 4.5 or higher on CIFAR-10 and 4

or higher on MNIST. On MNIST, the reported rate reaches 1.0 at $\epsilon_{\text{feat}} = 7.5$ and 10. These imply that aggregation can recover near-perfect agreement with noiseless predictions in the evaluated settings.

Third, PA improves utility across nearly all evaluated configurations. PrivUnit2+PA and PrivUnitG+PA generally provide the highest preservation rates and exhibit similar performance. At $K = 500$, $\epsilon_{\text{label}} = 3$, and $\epsilon_{\text{feat}} = 2$, PA increases the preservation rate of PrivUnit2 from 0.4520 to 0.7548 on CIFAR-10 and from 0.5480 to 0.8590 on MNIST. The corresponding improvements for PrivUnitG are from 0.4495 to 0.7612 and from 0.5650 to 0.8450. PA also substantially improves Laplace. At the same K and label budget, with $\epsilon_{\text{feat}} = 10$, the rate increases from 0.3033 to 0.9023 on CIFAR-10 and from 0.4250 to 0.9730 on MNIST. Across the evaluated base mechanisms, the few configurations in which PA does not improve preservation occur at small feature budgets, where preservation rates are low. The gains from PA also translate into lower feature budgets required to attain high preservation. At $K = 500$ and $\epsilon_{\text{label}} = 3$, both PA-enhanced PrivUnit mechanisms first exceed a preservation rate of 0.9 at $\epsilon_{\text{feat}} = 3$ on CIFAR-10 and $\epsilon_{\text{feat}} = 2.5$ on MNIST among the evaluated feature budgets.

Last, budget allocation matters at a fixed total privacy level. Table 17 compares allocations of the total per-item privacy budget $\epsilon_{\text{feat}} + \epsilon_{\text{label}}$ for PrivUnit2+PA at $K = 500$. For total budgets of 5 and 6, allocating more budget to labels improves preservation across the evaluated allocations. At a total budget of 6, the allocation $(\epsilon_{\text{feat}}, \epsilon_{\text{label}}) = (3, 3)$ achieves 0.9292 on CIFAR-10, compared with 0.7095 for $(4, 2)$ and 0.3450 for $(5, 1)$. The corresponding rates on MNIST are 0.9690, 0.8460, and 0.5005. However, at a total budget of 4, the balanced allocation $(2, 2)$ performs best on both datasets. This contrast illustrates the trade-off between reducing label perturbation and retaining sufficient budget for feature privatization. Among the allocations evaluated here, $\epsilon_{\text{label}} = 1$ consistently yields the lowest preservation at each fixed total budget.

Table 17: Prediction preservation rates of PrivUnit2+PA ($K = 500$) under fixed total privacy budgets $\epsilon_{\text{feat}} + \epsilon_{\text{label}}$. The highest rate among the evaluated allocations for each total budget and dataset is shown in bold.

Total ϵ	$(\epsilon_{\text{feat}}, \epsilon_{\text{label}})$	CIFAR-10	MNIST
4	$(3, 1)$	0.2213	0.3795
	$(2, 2)$	**0.4325**	**0.5720**
	$(1, 3)$	0.4025	0.4970
5	$(4, 1)$	0.2863	0.4375
	$(3, 2)$	0.6080	0.7480
	$(2, 3)$	**0.7548**	**0.8590**
6	$(5, 1)$	0.3450	0.5005
	$(4, 2)$	0.7095	0.8460
	$(3, 3)$	**0.9292**	**0.9690**

Table 18: Results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10 (ResNet-20)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 1.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.0618 (0.0193)	0.0630 (0.0185)	0.0630 (0.0188)	0.0640 (0.0186)	0.0645 (0.0182)	0.0658 (0.0184)	0.0668 (0.0184)	0.0698 (0.0180)	0.0708 (0.0183)	0.0713 (0.0176)	0.0783 (0.0184)	0.0843 (0.0211)
	200	0.0483 (0.0127)	0.0495 (0.0132)	0.0505 (0.0132)	0.0525 (0.0124)	0.0535 (0.0115)	0.0548 (0.0115)	0.0563 (0.0110)	0.0563 (0.0113)	0.0567 (0.0123)	0.0585 (0.0128)	0.0658 (0.0114)	0.0725 (0.0112)
	500	0.0463 (0.0150)	0.0473 (0.0146)	0.0488 (0.0145)	0.0508 (0.0150)	0.0520 (0.0157)	0.0533 (0.0158)	0.0538 (0.0144)	0.0560 (0.0146)	0.0573 (0.0146)	0.0600 (0.0151)	0.0693 (0.0204)	0.0788 (0.0232)
Laplace+PA	100	0.0645 (0.0172)	0.0680 (0.0179)	0.0720 (0.0170)	0.0788 (0.0165)	0.0845 (0.0159)	0.0915 (0.0168)	0.0968 (0.0171)	0.1030 (0.0181)	0.1090 (0.0194)	0.1155 (0.0186)	0.1498 (0.0205)	0.1872 (0.0239)
	200	0.0527 (0.0197)	0.0565 (0.0211)	0.0615 (0.0200)	0.0653 (0.0193)	0.0698 (0.0198)	0.0753 (0.0243)	0.0833 (0.0231)	0.0900 (0.0228)	0.0955 (0.0234)	0.1028 (0.0229)	0.1375 (0.0273)	0.1832 (0.0303)
	500	0.0485 (0.0118)	0.0553 (0.0129)	0.0588 (0.0120)	0.0648 (0.0128)	0.0748 (0.0167)	0.0823 (0.0179)	0.0953 (0.0185)	0.1067 (0.0181)	0.1188 (0.0183)	0.1278 (0.0187)	0.1883 (0.0283)	0.2602 (0.0284)
PrivUnit2	100	0.0700 (0.0210)	0.0820 (0.0194)	0.0895 (0.0195)	0.1028 (0.0234)	0.1153 (0.0224)	0.1273 (0.0206)	0.1378 (0.0245)	0.1510 (0.0240)	0.1600 (0.0218)	0.1685 (0.0250)	0.2238 (0.0280)	0.2890 (0.0283)
	200	0.0560 (0.0157)	0.0643 (0.0145)	0.0783 (0.0174)	0.0905 (0.0166)	0.1018 (0.0229)	0.1130 (0.0264)	0.1318 (0.0277)	0.1475 (0.0220)	0.1587 (0.0173)	0.1710 (0.0209)	0.2400 (0.0229)	0.3083 (0.0345)
	500	0.0513 (0.0163)	0.0615 (0.0177)	0.0762 (0.0201)	0.0955 (0.0229)	0.1170 (0.0242)	0.1430 (0.0274)	0.1710 (0.0276)	0.1868 (0.0275)	0.2182 (0.0363)	0.2457 (0.0316)	0.3533 (0.0256)	0.4487 (0.0326)
PrivUnit2+PA	100	0.0685 (0.0213)	0.0850 (0.0175)	0.1033 (0.0225)	0.1150 (0.0256)	0.1310 (0.0222)	0.1557 (0.0237)	0.1700 (0.0205)	0.1888 (0.0226)	0.2197 (0.0354)	0.2467 (0.0304)	0.3345 (0.0310)	0.3945 (0.0295)
	200	0.0622 (0.0192)	0.0788 (0.0190)	0.0980 (0.0244)	0.1210 (0.0230)	0.1440 (0.0308)	0.1690 (0.0309)	0.1945 (0.0294)	0.2145 (0.0280)	0.2400 (0.0301)	0.2628 (0.0316)	0.3425 (0.0278)	0.4115 (0.0344)
	500	0.0620 (0.0119)	0.0933 (0.0159)	0.1197 (0.0161)	0.1495 (0.0179)	0.1857 (0.0227)	0.2213 (0.0238)	0.2552 (0.0339)	0.2863 (0.0299)	0.3130 (0.0338)	0.3450 (0.0358)	0.4527 (0.0405)	0.5243 (0.0342)
PrivUnitG	100	0.0698 (0.0164)	0.0805 (0.0181)	0.0885 (0.0205)	0.1023 (0.0267)	0.1120 (0.0232)	0.1240 (0.0213)	0.1395 (0.0215)	0.1510 (0.0211)	0.1603 (0.0220)	0.1743 (0.0195)	0.2317 (0.0275)	0.2850 (0.0272)
	200	0.0520 (0.0165)	0.0595 (0.0170)	0.0727 (0.0215)	0.0845 (0.0221)	0.0958 (0.0228)	0.1087 (0.0226)	0.1247 (0.0235)	0.1395 (0.0218)	0.1520 (0.0229)	0.1720 (0.0306)	0.2467 (0.0326)	0.2997 (0.0332)
	500	0.0533 (0.0127)	0.0660 (0.0159)	0.0848 (0.0208)	0.1012 (0.0202)	0.1207 (0.0179)	0.1465 (0.0189)	0.1692 (0.0247)	0.1992 (0.0315)	0.2278 (0.0282)	0.2470 (0.0285)	0.3490 (0.0323)	0.4245 (0.0300)
PrivUnitG+PA	100	0.0690 (0.0166)	0.0898 (0.0171)	0.1045 (0.0174)	0.1220 (0.0186)	0.1430 (0.0209)	0.1630 (0.0255)	0.1815 (0.0250)	0.1970 (0.0250)	0.2180 (0.0240)	0.2397 (0.0289)	0.3145 (0.0338)	0.3737 (0.0295)
	200	0.0620 (0.0108)	0.0797 (0.0158)	0.0958 (0.0200)	0.1160 (0.0193)	0.1408 (0.0230)	0.1595 (0.0263)	0.1822 (0.0259)	0.2075 (0.0278)	0.2263 (0.0287)	0.2463 (0.0298)	0.3350 (0.0396)	0.3947 (0.0354)
	500	0.0550 (0.0195)	0.0735 (0.0235)	0.1015 (0.0265)	0.1295 (0.0236)	0.1620 (0.0299)	0.1990 (0.0319)	0.2395 (0.0340)	0.2770 (0.0325)	0.3063 (0.0332)	0.3362 (0.0349)	0.4460 (0.0416)	0.5115 (0.0430)

Table 19: Results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10 (ResNet-20)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 2.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.0853 (0.0234)	0.0870 (0.0239)	0.0890 (0.0240)	0.0898 (0.0240)	0.0917 (0.0245)	0.0938 (0.0253)	0.0958 (0.0256)	0.0975 (0.0256)	0.0998 (0.0260)	0.1020 (0.0269)	0.1115 (0.0274)	0.1217 (0.0296)
	200	0.0827 (0.0198)	0.0845 (0.0201)	0.0873 (0.0207)	0.0890 (0.0200)	0.0910 (0.0200)	0.0940 (0.0211)	0.0955 (0.0211)	0.0975 (0.0228)	0.1000 (0.0227)	0.1025 (0.0231)	0.1185 (0.0238)	0.1345 (0.0220)
	500	0.0858 (0.0231)	0.0912 (0.0230)	0.0935 (0.0222)	0.0980 (0.0231)	0.1022 (0.0236)	0.1048 (0.0229)	0.1075 (0.0230)	0.1088 (0.0225)	0.1155 (0.0237)	0.1207 (0.0239)	0.1462 (0.0275)	0.1755 (0.0294)
Laplace+PA	100	0.0800 (0.0193)	0.0870 (0.0210)	0.0930 (0.0216)	0.1032 (0.0215)	0.1088 (0.0227)	0.1195 (0.0234)	0.1310 (0.0213)	0.1418 (0.0225)	0.1518 (0.0226)	0.1610 (0.0227)	0.2248 (0.0328)	0.2898 (0.0308)
	200	0.0885 (0.0193)	0.0948 (0.0202)	0.1062 (0.0242)	0.1170 (0.0215)	0.1308 (0.0242)	0.1437 (0.0242)	0.1568 (0.0265)	0.1743 (0.0290)	0.1860 (0.0277)	0.1997 (0.0241)	0.2898 (0.0290)	0.3855 (0.0309)
	500	0.1000 (0.0244)	0.1195 (0.0276)	0.1388 (0.0271)	0.1620 (0.0270)	0.1837 (0.0315)	0.2080 (0.0351)	0.2350 (0.0346)	0.2630 (0.0406)	0.2920 (0.0397)	0.3175 (0.0400)	0.4740 (0.0360)	0.5938 (0.0378)
PrivUnit2	100	0.0927 (0.0161)	0.1082 (0.0173)	0.1295 (0.0156)	0.1535 (0.0188)	0.1742 (0.0173)	0.1942 (0.0186)	0.2180 (0.0231)	0.2350 (0.0259)	0.2603 (0.0295)	0.2773 (0.0357)	0.3775 (0.0281)	0.4530 (0.0455)
	200	0.1087 (0.0232)	0.1310 (0.0219)	0.1560 (0.0247)	0.1825 (0.0267)	0.2162 (0.0319)	0.2487 (0.0296)	0.2848 (0.0360)	0.3145 (0.0359)	0.3430 (0.0333)	0.3710 (0.0323)	0.5037 (0.0385)	0.5830 (0.0301)
	500	0.1140 (0.0245)	0.1580 (0.0293)	0.1998 (0.0314)	0.2530 (0.0329)	0.3040 (0.0362)	0.3660 (0.0363)	0.4163 (0.0359)	0.4710 (0.0446)	0.5132 (0.0369)	0.5520 (0.0342)	0.6923 (0.0366)	0.7550 (0.0354)
PrivUnit2+PA	100	0.1050 (0.0177)	0.1377 (0.0227)	0.1730 (0.0264)	0.2070 (0.0309)	0.2500 (0.0318)	0.2887 (0.0313)	0.3267 (0.0354)	0.3693 (0.0362)	0.4010 (0.0368)	0.4410 (0.0356)	0.5543 (0.0319)	0.6182 (0.0255)
	200	0.1215 (0.0222)	0.1598 (0.0185)	0.2140 (0.0219)	0.2677 (0.0272)	0.3265 (0.0314)	0.3820 (0.0330)	0.4350 (0.0336)	0.4840 (0.0344)	0.5315 (0.0363)	0.5658 (0.0310)	0.6805 (0.0269)	0.7345 (0.0337)
	500	0.1470 (0.0255)	0.2295 (0.0285)	0.3315 (0.0276)	0.4325 (0.0352)	0.5288 (0.0297)	0.6080 (0.0385)	0.6653 (0.0386)	0.7095 (0.0316)	0.7465 (0.0269)	0.7615 (0.0266)	0.8295 (0.0220)	0.8435 (0.0251)
PrivUnitG	100	0.0920 (0.0262)	0.1067 (0.0247)	0.1250 (0.0278)	0.1435 (0.0267)	0.1673 (0.0288)	0.1903 (0.0291)	0.2120 (0.0347)	0.2385 (0.0316)	0.2572 (0.0367)	0.2765 (0.0362)	0.3750 (0.0389)	0.4455 (0.0386)
	200	0.0965 (0.0211)	0.1183 (0.0234)	0.1438 (0.0217)	0.1725 (0.0241)	0.1988 (0.0262)	0.2273 (0.0354)	0.2640 (0.0366)	0.2918 (0.0429)	0.3243 (0.0388)	0.3598 (0.0417)	0.4872 (0.0428)	0.5670 (0.0422)
	500	0.1193 (0.0166)	0.1573 (0.0196)	0.2033 (0.0221)	0.2562 (0.0261)	0.3095 (0.0316)	0.3597 (0.0357)	0.4105 (0.0348)	0.4640 (0.0333)	0.5105 (0.0304)	0.5470 (0.0291)	0.6788 (0.0262)	0.7378 (0.0302)
PrivUnitG+PA	100	0.1070 (0.0235)	0.1365 (0.0280)	0.1732 (0.0300)	0.2055 (0.0256)	0.2480 (0.0292)	0.2840 (0.0280)	0.3220 (0.0303)	0.3533 (0.0299)	0.3910 (0.0297)	0.4195 (0.0308)	0.5260 (0.0355)	0.5998 (0.0298)
	200	0.1245 (0.0203)	0.1733 (0.0223)	0.2270 (0.0262)	0.2818 (0.0329)	0.3407 (0.0354)	0.3855 (0.0367)	0.4318 (0.0360)	0.4790 (0.0414)	0.5268 (0.0407)	0.5548 (0.0361)	0.6685 (0.0354)	0.7345 (0.0337)
	500	0.1380 (0.0257)	0.2157 (0.0343)	0.3202 (0.0348)	0.4255 (0.0391)	0.5198 (0.0428)	0.5913 (0.0425)	0.6512 (0.0413)	0.6977 (0.0411)	0.7308 (0.0366)	0.7585 (0.0310)	0.8200 (0.0264)	0.8430 (0.0266)

Table 20: Results for **Nonlinear MLP Classifier with ReLU** ($m = 64$) on **CIFAR-10 (ResNet-20)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 3.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.1048 (0.0170)	0.1083 (0.0170)	0.1108 (0.0173)	0.1140 (0.0174)	0.1170 (0.0179)	0.1213 (0.0181)	0.1247 (0.0191)	0.1290 (0.0206)	0.1323 (0.0215)	0.1365 (0.0228)	0.1585 (0.0229)	0.1782 (0.0253)
	200	0.1010 (0.0152)	0.1052 (0.0159)	0.1105 (0.0183)	0.1143 (0.0175)	0.1182 (0.0175)	0.1227 (0.0188)	0.1280 (0.0183)	0.1338 (0.0178)	0.1395 (0.0188)	0.1452 (0.0200)	0.1750 (0.0205)	0.2075 (0.0231)
	500	0.1040 (0.0205)	0.1103 (0.0196)	0.1163 (0.0195)	0.1253 (0.0186)	0.1338 (0.0208)	0.1428 (0.0236)	0.1518 (0.0260)	0.1613 (0.0265)	0.1710 (0.0257)	0.1810 (0.0255)	0.2393 (0.0290)	0.3033 (0.0340)
Laplace+PA	100	0.1042 (0.0169)	0.1170 (0.0179)	0.1293 (0.0189)	0.1420 (0.0214)	0.1557 (0.0239)	0.1760 (0.0231)	0.1945 (0.0233)	0.2112 (0.0289)	0.2308 (0.0271)	0.2562 (0.0244)	0.3617 (0.0225)	0.4797 (0.0335)
	200	0.1103 (0.0222)	0.1293 (0.0232)	0.1500 (0.0234)	0.1723 (0.0277)	0.1950 (0.0290)	0.2223 (0.0300)	0.2512 (0.0297)	0.2800 (0.0364)	0.3145 (0.0384)	0.3460 (0.0390)	0.5150 (0.0377)	0.6680 (0.0396)
	500	0.1215 (0.0233)	0.1547 (0.0290)	0.1970 (0.0337)	0.2372 (0.0400)	0.2847 (0.0407)	0.3350 (0.0458)	0.3957 (0.0461)	0.4557 (0.0535)	0.5145 (0.0477)	0.5677 (0.0442)	0.7870 (0.0283)	0.9023 (0.0224)
PrivUnit2	100	0.1308 (0.0213)	0.1575 (0.0240)	0.1895 (0.0217)	0.2215 (0.0250)	0.2598 (0.0273)	0.3020 (0.0242)	0.3520 (0.0344)	0.3955 (0.0302)	0.4300 (0.0267)	0.4642 (0.0379)	0.6175 (0.0370)	0.7210 (0.0352)
	200	0.1373 (0.0192)	0.1892 (0.0225)	0.2453 (0.0294)	0.3005 (0.0384)	0.3563 (0.0401)	0.4145 (0.0475)	0.4787 (0.0493)	0.5370 (0.0478)	0.5940 (0.0396)	0.6400 (0.0314)	0.8180 (0.0287)	0.8918 (0.0173)
	500	0.1593 (0.0237)	0.2405 (0.0306)	0.3360 (0.0344)	0.4520 (0.0245)	0.5565 (0.0240)	0.6650 (0.0291)	0.7535 (0.0294)	0.8035 (0.0284)	0.8510 (0.0301)	0.8815 (0.0296)	0.9650 (0.0105)	0.9855 (0.0076)
PrivUnit2+PA	100	0.1410 (0.0202)	0.2002 (0.0231)	0.2730 (0.0234)	0.3415 (0.0274)	0.4185 (0.0291)	0.4880 (0.0296)	0.5575 (0.0327)	0.6185 (0.0325)	0.6700 (0.0316)	0.7155 (0.0342)	0.8220 (0.0293)	0.8775 (0.0223)
	200	0.1700 (0.0228)	0.2677 (0.0301)	0.3760 (0.0307)	0.4928 (0.0272)	0.5982 (0.0351)	0.6873 (0.0339)	0.7632 (0.0313)	0.8117 (0.0243)	0.8593 (0.0247)	0.8870 (0.0240)	0.9560 (0.0149)	0.9735 (0.0114)
	500	0.2248 (0.0252)	0.4025 (0.0287)	0.6065 (0.0385)	0.7548 (0.0310)	0.8565 (0.0166)	0.9292 (0.0174)	0.9615 (0.0139)	0.9783 (0.0110)	0.9900 (0.0065)	0.9920 (0.0055)	0.9988 (0.0022)	0.9995 (0.0015)
PrivUnitG	100	0.1210 (0.0204)	0.1433 (0.0237)	0.1772 (0.0232)	0.2080 (0.0294)	0.2452 (0.0284)	0.2858 (0.0261)	0.3257 (0.0318)	0.3782 (0.0288)	0.4205 (0.0287)	0.4583 (0.0288)	0.6230 (0.0305)	0.7145 (0.0227)
	200	0.1360 (0.0219)	0.1797 (0.0229)	0.2298 (0.0253)	0.2838 (0.0296)	0.3440 (0.0355)	0.4055 (0.0298)	0.4787 (0.0312)	0.5292 (0.0306)	0.5883 (0.0327)	0.6370 (0.0272)	0.8002 (0.0243)	0.8757 (0.0273)
	500	0.1583 (0.0274)	0.2442 (0.0346)	0.3452 (0.0345)	0.4495 (0.0300)	0.5528 (0.0286)	0.6508 (0.0270)	0.7345 (0.0238)	0.7970 (0.0193)	0.8450 (0.0193)	0.8828 (0.0182)	0.9620 (0.0112)	0.9845 (0.0071)
PrivUnitG+PA	100	0.1338 (0.0241)	0.1972 (0.0310)	0.2633 (0.0333)	0.3430 (0.0374)	0.4135 (0.0423)	0.4905 (0.0364)	0.5578 (0.0374)	0.6140 (0.0361)	0.6615 (0.0369)	0.7038 (0.0433)	0.8223 (0.0284)	0.8783 (0.0253)
	200	0.1785 (0.0229)	0.2685 (0.0257)	0.3790 (0.0341)	0.5007 (0.0308)	0.6035 (0.0257)	0.6945 (0.0252)	0.7682 (0.0335)	0.8207 (0.0235)	0.8608 (0.0197)	0.8880 (0.0180)	0.9532 (0.0138)	0.9715 (0.0130)
	500	0.2185 (0.0240)	0.4027 (0.0275)	0.6030 (0.0319)	0.7612 (0.0306)	0.8628 (0.0200)	0.9255 (0.0169)	0.9573 (0.0104)	0.9758 (0.0080)	0.9850 (0.0061)	0.9905 (0.0063)	0.9985 (0.0029)	0.9988 (0.0022)

Table 21: Results for **Nonlinear MLP Classifier with ReLU** ($m = 32$) on **MNIST (VAE)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 1.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.0950 (0.0155)	0.0955 (0.0159)	0.0980 (0.0165)	0.0985 (0.0168)	0.1005 (0.0150)	0.1015 (0.0142)	0.1020 (0.0149)	0.1030 (0.0157)	0.1040 (0.0154)	0.1070 (0.0169)	0.1165 (0.0219)	0.1300 (0.0267)
	200	0.0925 (0.0144)	0.0945 (0.0150)	0.0985 (0.0147)	0.1025 (0.0157)	0.1040 (0.0154)	0.1055 (0.0155)	0.1075 (0.0174)	0.1090 (0.0168)	0.1085 (0.0163)	0.1095 (0.0172)	0.1175 (0.0195)	0.1265 (0.0196)
	500	0.0985 (0.0244)	0.1015 (0.0252)	0.1015 (0.0257)	0.1040 (0.0257)	0.1075 (0.0237)	0.1120 (0.0241)	0.1130 (0.0244)	0.1170 (0.0249)	0.1205 (0.0215)	0.1235 (0.0199)	0.1425 (0.0232)	0.1660 (0.0294)
Laplace+PA	100	0.1065 (0.0255)	0.1105 (0.0273)	0.1165 (0.0261)	0.1275 (0.0215)	0.1380 (0.0232)	0.1420 (0.0223)	0.1510 (0.0254)	0.1615 (0.0211)	0.1695 (0.0155)	0.1770 (0.0134)	0.2350 (0.0143)	0.2865 (0.0136)
	200	0.0970 (0.0220)	0.1035 (0.0201)	0.1105 (0.0185)	0.1120 (0.0196)	0.1175 (0.0216)	0.1305 (0.0211)	0.1485 (0.0284)	0.1580 (0.0316)	0.1695 (0.0317)	0.1780 (0.0377)	0.2295 (0.0277)	0.2935 (0.0243)
	500	0.1055 (0.0230)	0.1135 (0.0275)	0.1255 (0.0297)	0.1370 (0.0256)	0.1490 (0.0271)	0.1690 (0.0275)	0.1845 (0.0294)	0.2020 (0.0366)	0.2140 (0.0384)	0.2285 (0.0397)	0.2990 (0.0365)	0.3660 (0.0323)
PrivUnit2	100	0.1235 (0.0176)	0.1390 (0.0131)	0.1475 (0.0134)	0.1640 (0.0154)	0.1880 (0.0230)	0.2060 (0.0161)	0.2170 (0.0231)	0.2465 (0.0292)	0.2550 (0.0186)	0.2625 (0.0307)	0.3310 (0.0357)	0.4165 (0.0376)
	200	0.1085 (0.0206)	0.1205 (0.0161)	0.1395 (0.0174)	0.1590 (0.0202)	0.1810 (0.0211)	0.2050 (0.0260)	0.2165 (0.0264)	0.2360 (0.0289)	0.2675 (0.0323)	0.2860 (0.0221)	0.3555 (0.0250)	0.4360 (0.0378)
	500	0.1090 (0.0170)	0.1220 (0.0196)	0.1430 (0.0155)	0.1715 (0.0227)	0.1995 (0.0224)	0.2210 (0.0229)	0.2560 (0.0293)	0.2685 (0.0370)	0.2880 (0.0431)	0.3160 (0.0278)	0.4220 (0.0300)	0.4970 (0.0186)
PrivUnit2+PA	100	0.1145 (0.0130)	0.1365 (0.0219)	0.1715 (0.0254)	0.2005 (0.0199)	0.2320 (0.0288)	0.2670 (0.0221)	0.3015 (0.0193)	0.3390 (0.0281)	0.3535 (0.0167)	0.3900 (0.0224)	0.4755 (0.0432)	0.5675 (0.0444)
	200	0.1355 (0.0188)	0.1635 (0.0207)	0.1975 (0.0261)	0.2350 (0.0312)	0.2685 (0.0233)	0.3065 (0.0226)	0.3405 (0.0326)	0.3675 (0.0333)	0.3930 (0.0357)	0.4340 (0.0383)	0.5360 (0.0414)	0.6065 (0.0349)
	500	0.1460 (0.0251)	0.1845 (0.0272)	0.2280 (0.0246)	0.2715 (0.0422)	0.3315 (0.0378)	0.3795 (0.0290)	0.4080 (0.0334)	0.4375 (0.0335)	0.4675 (0.0353)	0.5005 (0.0377)	0.5925 (0.0325)	0.6620 (0.0210)
PrivUnitG	100	0.1160 (0.0196)	0.1265 (0.0224)	0.1460 (0.0301)	0.1550 (0.0312)	0.1675 (0.0299)	0.1890 (0.0288)	0.2090 (0.0311)	0.2240 (0.0338)	0.2355 (0.0299)	0.2500 (0.0332)	0.3185 (0.0295)	0.3670 (0.0333)
	200	0.1030 (0.0158)	0.1160 (0.0147)	0.1375 (0.0132)	0.1525 (0.0136)	0.1695 (0.0138)	0.1925 (0.0210)	0.2130 (0.0189)	0.2335 (0.0180)	0.2495 (0.0196)	0.2715 (0.0243)	0.3510 (0.0238)	0.4110 (0.0286)
	500	0.1205 (0.0161)	0.1420 (0.0226)	0.1620 (0.0178)	0.1915 (0.0315)	0.2215 (0.0277)	0.2430 (0.0363)	0.2685 (0.0410)	0.2925 (0.0340)	0.3165 (0.0419)	0.3400 (0.0437)	0.4290 (0.0431)	0.4925 (0.0392)
PrivUnitG+PA	100	0.1210 (0.0232)	0.1460 (0.0204)	0.1715 (0.0246)	0.2015 (0.0272)	0.2320 (0.0323)	0.2630 (0.0341)	0.2880 (0.0284)	0.3135 (0.0304)	0.3435 (0.0257)	0.3710 (0.0231)	0.4800 (0.0268)	0.5450 (0.0268)
	200	0.1210 (0.0232)	0.1505 (0.0249)	0.1885 (0.0184)	0.2210 (0.0188)	0.2570 (0.0228)	0.2930 (0.0304)	0.3325 (0.0286)	0.3700 (0.0374)	0.3940 (0.0418)	0.4150 (0.0395)	0.5175 (0.0452)	0.5720 (0.0454)
	500	0.1305 (0.0194)	0.1715 (0.0283)	0.2125 (0.0302)	0.2655 (0.0326)	0.3130 (0.0345)	0.3495 (0.0289)	0.3955 (0.0312)	0.4385 (0.0247)	0.4755 (0.0337)	0.4995 (0.0332)	0.5940 (0.0327)	0.6490 (0.0303)

Table 22: Results for **Nonlinear MLP Classifier with ReLU** ($m = 32$) on **MNIST (VAE)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 2.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.0945 (0.0172)	0.0955 (0.0167)	0.1000 (0.0163)	0.1020 (0.0169)	0.1095 (0.0148)	0.1110 (0.0154)	0.1145 (0.0177)	0.1195 (0.0189)	0.1235 (0.0197)	0.1270 (0.0206)	0.1420 (0.0208)	0.1670 (0.0207)
	200	0.0915 (0.0236)	0.0975 (0.0262)	0.1080 (0.0278)	0.1115 (0.0283)	0.1155 (0.0296)	0.1210 (0.0266)	0.1230 (0.0236)	0.1285 (0.0233)	0.1330 (0.0216)	0.1350 (0.0209)	0.1590 (0.0255)	0.1875 (0.0235)
	500	0.1050 (0.0120)	0.1090 (0.0124)	0.1140 (0.0145)	0.1185 (0.0142)	0.1260 (0.0174)	0.1400 (0.0163)	0.1475 (0.0190)	0.1530 (0.0208)	0.1650 (0.0212)	0.1695 (0.0218)	0.2125 (0.0235)	0.2675 (0.0289)
Laplace+PA	100	0.1125 (0.0215)	0.1260 (0.0244)	0.1365 (0.0246)	0.1505 (0.0285)	0.1650 (0.0291)	0.1780 (0.0274)	0.1915 (0.0293)	0.2095 (0.0308)	0.2300 (0.0295)	0.2450 (0.0316)	0.3260 (0.0281)	0.4160 (0.0328)
	200	0.1075 (0.0127)	0.1180 (0.0160)	0.1360 (0.0201)	0.1530 (0.0270)	0.1685 (0.0264)	0.1880 (0.0251)	0.2125 (0.0244)	0.2445 (0.0259)	0.2655 (0.0177)	0.2875 (0.0193)	0.4035 (0.0327)	0.5115 (0.0341)
	500	0.1290 (0.0336)	0.1435 (0.0302)	0.1775 (0.0327)	0.2115 (0.0350)	0.2480 (0.0288)	0.2835 (0.0304)	0.3260 (0.0325)	0.3640 (0.0345)	0.4000 (0.0367)	0.4395 (0.0349)	0.6085 (0.0418)	0.7360 (0.0299)
PrivUnit2	100	0.1225 (0.0172)	0.1400 (0.0208)	0.1615 (0.0219)	0.1880 (0.0275)	0.2205 (0.0182)	0.2540 (0.0190)	0.2770 (0.0190)	0.3030 (0.0322)	0.3390 (0.0339)	0.3620 (0.0275)	0.4925 (0.0372)	0.5870 (0.0300)
	200	0.1220 (0.0177)	0.1515 (0.0178)	0.1900 (0.0212)	0.2305 (0.0263)	0.2735 (0.0285)	0.3155 (0.0314)	0.3640 (0.0303)	0.3955 (0.0280)	0.4415 (0.0317)	0.4705 (0.0293)	0.6290 (0.0431)	0.7260 (0.0379)
	500	0.1275 (0.0251)	0.1860 (0.0246)	0.2545 (0.0353)	0.3150 (0.0288)	0.3980 (0.0311)	0.4575 (0.0255)	0.5200 (0.0416)	0.5805 (0.0309)	0.6320 (0.0427)	0.6715 (0.0372)	0.8140 (0.0275)	0.9050 (0.0231)
PrivUnit2+PA	100	0.1285 (0.0243)	0.1740 (0.0273)	0.2210 (0.0213)	0.2860 (0.0313)	0.3345 (0.0368)	0.3855 (0.0322)	0.4310 (0.0406)	0.4945 (0.0488)	0.5375 (0.0301)	0.5590 (0.0387)	0.6875 (0.0307)	0.7560 (0.0209)
	200	0.1545 (0.0320)	0.2220 (0.0350)	0.3090 (0.0311)	0.3805 (0.0377)	0.4620 (0.0223)	0.5370 (0.0215)	0.5925 (0.0207)	0.6470 (0.0335)	0.6855 (0.0293)	0.7105 (0.0194)	0.8225 (0.0164)	0.8690 (0.0207)
	500	0.1835 (0.0245)	0.3060 (0.0275)	0.4440 (0.0237)	0.5720 (0.0271)	0.6730 (0.0334)	0.7480 (0.0334)	0.8010 (0.0329)	0.8460 (0.0197)	0.8690 (0.0213)	0.8875 (0.0215)	0.9490 (0.0229)	0.9655 (0.0107)
PrivUnitG	100	0.1190 (0.0309)	0.1390 (0.0326)	0.1615 (0.0292)	0.1895 (0.0291)	0.2110 (0.0270)	0.2355 (0.0218)	0.2665 (0.0275)	0.2955 (0.0230)	0.3180 (0.0220)	0.3475 (0.0247)	0.4525 (0.0270)	0.5345 (0.0318)
	200	0.1190 (0.0251)	0.1590 (0.0332)	0.2020 (0.0383)	0.2380 (0.0365)	0.2790 (0.0350)	0.3175 (0.0352)	0.3575 (0.0355)	0.3975 (0.0351)	0.4345 (0.0332)	0.4735 (0.0349)	0.6095 (0.0372)	0.6860 (0.0181)
	500	0.1385 (0.0162)	0.2000 (0.0216)	0.2675 (0.0332)	0.3380 (0.0367)	0.4065 (0.0354)	0.4810 (0.0320)	0.5395 (0.0346)	0.5890 (0.0290)	0.6330 (0.0299)	0.6810 (0.0237)	0.8295 (0.0107)	0.8835 (0.0155)
PrivUnitG+PA	100	0.1225 (0.0246)	0.1585 (0.0305)	0.2060 (0.0248)	0.2690 (0.0368)	0.3175 (0.0331)	0.3695 (0.0343)	0.4240 (0.0440)	0.4670 (0.0304)	0.5075 (0.0264)	0.5505 (0.0281)	0.6665 (0.0370)	0.7320 (0.0365)
	200	0.1430 (0.0306)	0.2200 (0.0307)	0.2910 (0.0271)	0.3695 (0.0214)	0.4440 (0.0240)	0.5140 (0.0279)	0.5730 (0.0263)	0.6350 (0.0270)	0.6745 (0.0210)	0.7120 (0.0243)	0.8095 (0.0253)	0.8620 (0.0251)
	500	0.1860 (0.0259)	0.3050 (0.0290)	0.4355 (0.0323)	0.5615 (0.0296)	0.6565 (0.0346)	0.7335 (0.0183)	0.7830 (0.0228)	0.8270 (0.0246)	0.8600 (0.0262)	0.8820 (0.0255)	0.9390 (0.0143)	0.9530 (0.0170)

Table 23: Results for **Nonlinear MLP Classifier with ReLU** ($m = 32$) on **MNIST (VAE)** under ϵ-LDP mechanisms. Label-Anchored $\epsilon_{\text{label}} = 3.0$ (fixed).

		Privacy Budget (ϵ_{feat})											
Mechanism	K	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	7.5	10.0
Laplace (ℓ_1)	100	0.1030 (0.0266)	0.1075 (0.0264)	0.1130 (0.0261)	0.1160 (0.0248)	0.1195 (0.0272)	0.1265 (0.0267)	0.1340 (0.0238)	0.1385 (0.0246)	0.1465 (0.0281)	0.1495 (0.0275)	0.1815 (0.0285)	0.2210 (0.0312)
	200	0.1045 (0.0183)	0.1100 (0.0225)	0.1140 (0.0218)	0.1195 (0.0217)	0.1255 (0.0213)	0.1335 (0.0239)	0.1375 (0.0200)	0.1440 (0.0228)	0.1515 (0.0224)	0.1625 (0.0251)	0.2145 (0.0224)	0.2710 (0.0225)
	500	0.1050 (0.0226)	0.1140 (0.0227)	0.1265 (0.0253)	0.1370 (0.0266)	0.1480 (0.0218)	0.1635 (0.0220)	0.1780 (0.0212)	0.1935 (0.0215)	0.2115 (0.0229)	0.2305 (0.0283)	0.3245 (0.0428)	0.4250 (0.0397)
Laplace+PA	100	0.1115 (0.0239)	0.1320 (0.0239)	0.1610 (0.0320)	0.1840 (0.0394)	0.2110 (0.0369)	0.2370 (0.0342)	0.2640 (0.0319)	0.2960 (0.0294)	0.3260 (0.0313)	0.3525 (0.0316)	0.5110 (0.0384)	0.6385 (0.0283)
	200	0.1240 (0.0245)	0.1470 (0.0297)	0.1770 (0.0339)	0.2065 (0.0355)	0.2400 (0.0370)	0.2810 (0.0397)	0.3325 (0.0372)	0.3735 (0.0316)	0.4180 (0.0347)	0.4665 (0.0356)	0.6690 (0.0228)	0.8110 (0.0253)
	500	0.1420 (0.0181)	0.1855 (0.0186)	0.2425 (0.0338)	0.3025 (0.0338)	0.3660 (0.0294)	0.4495 (0.0358)	0.5205 (0.0394)	0.5900 (0.0264)	0.6655 (0.0222)	0.7170 (0.0212)	0.9090 (0.0145)	0.9730 (0.0089)
PrivUnit2	100	0.1270 (0.0281)	0.1665 (0.0292)	0.2090 (0.0315)	0.2560 (0.0294)	0.3150 (0.0346)	0.3630 (0.0369)	0.4165 (0.0347)	0.4810 (0.0479)	0.5335 (0.0428)	0.5585 (0.0442)	0.7425 (0.0218)	0.8275 (0.0216)
	200	0.1405 (0.0177)	0.1935 (0.0357)	0.2830 (0.0314)	0.3615 (0.0299)	0.4515 (0.0362)	0.5345 (0.0364)	0.6015 (0.0273)	0.6635 (0.0382)	0.7095 (0.0277)	0.7555 (0.0311)	0.8965 (0.0235)	0.9510 (0.0139)
	500	0.1655 (0.0305)	0.2880 (0.0317)	0.4155 (0.0411)	0.5480 (0.0360)	0.6765 (0.0329)	0.7720 (0.0351)	0.8455 (0.0240)	0.8975 (0.0210)	0.9285 (0.0204)	0.9500 (0.0103)	0.9920 (0.0067)	0.9980 (0.0035)
PrivUnit2+PA	100	0.1530 (0.0261)	0.2300 (0.0238)	0.3195 (0.0408)	0.4120 (0.0391)	0.5260 (0.0517)	0.5990 (0.0498)	0.6740 (0.0482)	0.7400 (0.0416)	0.7860 (0.0360)	0.8215 (0.0364)	0.8935 (0.0129)	0.9400 (0.0158)
	200	0.1900 (0.0163)	0.3195 (0.0229)	0.4660 (0.0301)	0.5935 (0.0265)	0.7215 (0.0176)	0.8020 (0.0155)	0.8585 (0.0181)	0.8985 (0.0215)	0.9245 (0.0172)	0.9420 (0.0157)	0.9785 (0.0100)	0.9890 (0.0074)
	500	0.2570 (0.0258)	0.4970 (0.0274)	0.7160 (0.0245)	0.8590 (0.0117)	0.9355 (0.0093)	0.9690 (0.0133)	0.9840 (0.0105)	0.9925 (0.0068)	0.9950 (0.0062)	0.9970 (0.0042)	1.0000 (0.0000)	1.0000 (0.0000)
PrivUnitG	100	0.1280 (0.0225)	0.1650 (0.0362)	0.2120 (0.0387)	0.2660 (0.0429)	0.3175 (0.0428)	0.3640 (0.0386)	0.4150 (0.0318)	0.4655 (0.0335)	0.5170 (0.0358)	0.5560 (0.0352)	0.7080 (0.0208)	0.7935 (0.0186)
	200	0.1320 (0.0208)	0.1905 (0.0179)	0.2705 (0.0172)	0.3555 (0.0271)	0.4280 (0.0307)	0.5035 (0.0290)	0.5780 (0.0252)	0.6425 (0.0355)	0.6940 (0.0341)	0.7420 (0.0324)	0.8805 (0.0188)	0.9385 (0.0116)
	500	0.1800 (0.0183)	0.3000 (0.0203)	0.4280 (0.0207)	0.5650 (0.0190)	0.6780 (0.0406)	0.7665 (0.0406)	0.8360 (0.0333)	0.8875 (0.0269)	0.9165 (0.0266)	0.9435 (0.0131)	0.9925 (0.0075)	0.9990 (0.0021)
PrivUnitG+PA	100	0.1510 (0.0251)	0.2255 (0.0271)	0.3080 (0.0272)	0.4110 (0.0257)	0.5110 (0.0353)	0.5925 (0.0434)	0.6685 (0.0323)	0.7250 (0.0342)	0.7735 (0.0358)	0.8055 (0.0353)	0.8940 (0.0231)	0.9290 (0.0185)
	200	0.1815 (0.0197)	0.3115 (0.0292)	0.4485 (0.0286)	0.5895 (0.0239)	0.7005 (0.0251)	0.7960 (0.0298)	0.8625 (0.0276)	0.9040 (0.0168)	0.9255 (0.0164)	0.9450 (0.0143)	0.9815 (0.0078)	0.9915 (0.0024)
	500	0.2525 (0.0357)	0.4795 (0.0192)	0.7040 (0.0260)	0.8450 (0.0264)	0.9215 (0.0227)	0.9655 (0.0117)	0.9840 (0.0097)	0.9920 (0.0089)	0.9950 (0.0078)	0.9985 (0.0034)	1.0000 (0.0000)	1.0000 (0.0000)