PAVE Specifications of Learnwares Yield Intrinsic Privacy-Preserving Capabilities

Hao-Yi Lei[1,2] , **Jin-Hui Wu**[3] , **Zhi-Hao Tan**[1,2] , **Zhi-Hua Zhou**[1,2]

[1] National Key Laboratory for Novel Software Technology, Nanjing University, China
[2] School of Artificial Intelligence, Nanjing University, China
[3] School of Computer Science and Technology, Soochow University, China
{leihy, tanzh, zhouzh}@lamda.nju.edu.cn, wujinhui@suda.edu.cn,

Abstract

The learnware paradigm supports model reuse by pairing each submitted model with a *specification*, a lightweight representation used by the learnware dock system to identify, match, and reuse models without accessing raw data. While specifications are essential for learnware identification, they are also data-dependent public artifacts and it is not clear whether they reveal private information. Recently, the *Parameter Vector* (PAVE) specification has been proposed and shown to be effective for learnwares, yet its privacy properties remain largely unexplored. In this paper, we provide the first theoretical privacy analysis for PAVE. Specifically, we first formalize two specification-induced risks in the learnware paradigm: the *disclosure risk* of the released specification and the *amplification risk* that the specification may strengthen attacks against the released model. Second, we characterize when compact PAVE releases admit intrinsic differential privacy: under natural structural conditions of learnware docks, the compact PAVE specification satisfies an (ε, δ)-DP guarantee without explicit additive noise through a Gaussian-sketch view of stable parameter variations, and for regimes outside these conditions, we further provide DP-S-PAVE as a certified differentially private variant. Third, we show that the resulting DP guarantees control both disclosure risk and specification-side amplification risk, and we analyze the induced privacy–utility trade-off to guide effective learnware identification while preserving privacy.

1 Introduction

Reusing existing models is often more practical than training a new model from scratch. In many applications, developers already hold trained models that may be useful for future tasks, while users often have only limited data and limited computational resources. The difficulty is that model reuse is not only a matter of collecting models. A user must first identify which existing models are relevant to her task. Simple metadata, such as model names, architectures, tags, or benchmark scores, can only provide a rough description of model capability. Directly evaluating many candidate models on the user's raw data is more informative, but it is costly and may expose private task information. Asking developers to reveal their training data would also help model identification, but is usually infeasible due to privacy or proprietary concerns.

The learnware paradigm [Zhou, 2016; Zhou and Tan, 2024] addresses this problem by making the specification an explicit part of model reuse. A learnware is not only a trained model, but a model accompanied by a specification that describes its reusable capability. Such a specification can be generated from the model, the training data, or task information, and the dock system can use it to match developer-side models with user-side requirements without raw-data exchange. In this sense, the specification is the public artifact that makes learnware identification possible. However, this specification is also data-dependent. The more accurately it reflects model capability, the more likely it is to carry information about the data or task from which it was generated. Thus, the privacy question in learnware is not only about releasing models, but also about releasing specifications: what information does this public capability specification reveal, and how can such leakage be controlled while keeping the specification useful?

Recently, PAVE, a parameter-vector specification for learnwares, has shown strong empirical effectiveness for learnware identification and reuse [Shi *et al.*, 2026]. Unlike reduced-set specifications such as RKME-style specifications [Wu *et al.*, 2023], PAVE summarizes task and capability information through parameter variations induced by fine-tuning. The released specification is a compact low-rank parameter object rather than a synthetic representative set in the data space. This difference makes existing privacy analyses for reduced-set specifications insufficient for PAVE. On the one hand, gradient-induced parameter variations may themselves carry information about the underlying data. On the other hand, a PAVE specification is released together with a model, so it may also provide additional evidence for attacks against the released model. These observations motivate a privacy analysis tailored to the PAVE generation and release procedure.

Differential privacy (DP) [Dwork, 2006] is a natural tool for such an analysis, since it provides a robust worst-case

A complete version with detailed proofs is available at https://default-anno-bucket.s3.us-west-1.amazonaws.com/Proof.pdf.

guarantee against inference attacks. However, applying DP to learnware specifications is non-trivial. Most DP mechanisms rely on explicit noise injection, while learnware specifications are intended to be compact and useful for identification. Excessive perturbation can therefore damage the matching utility of the learnware dock system, as also observed in prior studies on learnware privacy [Lei *et al.*, 2024]. Moreover, a specification-level privacy guarantee should be connected to learnware-specific risks: releasing a specification may leak information by itself, and may also strengthen attacks when combined with the released model.

In this work, we characterize the privacy-preserving capabilities of PAVE specifications. We analyze the compact PAVE release used for identification, study structural conditions under which it admits intrinsic DP, and complement this characterization with DP-S-PAVE for settings where these conditions are not enforced. We then study how the resulting DP guarantees translate into learnware-level risk guarantees. The main contributions are summarized as follows:

- We formalize two specification-induced privacy risks for PAVE, namely the disclosure risk of the released specification and the amplification risk that the specification may strengthen attacks against the released model.
- We show that compact PAVE releases admit an intrinsic (ε, δ)-DP guarantee under natural structural conditions of learnware docks, through a Gaussian-sketch view of stable parameter variations. For regimes outside these conditions, we further provide DP-S-PAVE as a certified differentially private variant.
- We prove that the resulting DP guarantees control both disclosure risk and specification-side amplification risk, and analyze the induced privacy–utility trade-off to guide privacy-aware deployment while preserving effective learnware identification.

2 Preliminaries and Methodology

In this section, we introduce the basic workflow of the learnware paradigm with PAVE specifications, and formalize the privacy concerns that arise from releasing such specifications.

2.1 Background of Learnware Paradigm

We first recall the two-stage learnware workflow.

Submitting stage. For a developer, a learnware consists of a trained model h together with its specification. PAVE constructs a *model vector* τ_h by fine-tuning a shared pre-trained model $f(\cdot;\theta_0)$ on the developer's training dataset D_t with a task-specific loss L_t. When the output space of the pre-trained model differs from that of the task, a lightweight mapping g_t is used to bridge the two spaces. Concretely, PAVE is obtained as the parameter update that minimizes

$$\tau_h \;=\; \arg\min_{\tau} \sum_{(x,y)\in D_t} L_t\big(g_t \circ f(x;\theta_0+\tau),\, h(x)\big). \quad (1)$$

The learnware (h, τ_h) is then submitted to the dock system.

Deploying stage. Given a user task with a few-shot dataset D_u and loss L_u, PAVE constructs a *task vector* τ_u using the same pre-trained model $f(\cdot;\theta_0)$, but replaces the model prediction target by the ground-truth label. Thus the vector reflects the capability required by the user task:

$$\tau_u \;=\; \arg\min_{\tau} \sum_{(x,y)\in D_u} L_u\big(g_u \circ f(x;\theta_0+\tau),\, y\big). \quad (2)$$

The dock system identifies helpful learnwares by comparing cosine similarity in the parameter-vector space:

$$\cos(\tau_h, \tau_u) \;=\; \frac{\langle \tau_h, \tau_u \rangle}{\|\tau_h\|_2 \, \|\tau_u\|_2}, \quad (3)$$

and selects learnwares with large similarity for reuse.

To make PAVE efficient to store and compare, it adopts a LoRA-style low-rank parameterization. For each selected weight matrix $\mathbf{W}_\ell$ in the pre-trained backbone, the update is represented as

$$\Delta \mathbf{W}_\ell = \mathbf{B}_\ell \mathbf{A}_\ell, \quad (4)$$

where $\mathbf{A}_\ell \in \mathbb{R}^{r\times n_\ell}$ is randomly initialized and then frozen, while $\mathbf{B}_\ell \in \mathbb{R}^{m_\ell \times r}$ is learned from data, with $r \ll \min\{m_\ell, n_\ell\}$. Writing $\mathbf{A}$ for the collection of frozen factors, PAVE learns the compact factors by

$$\mathbf{B}^\star = \arg\min_{\mathbf{B}} \sum_{(x,y)\in D} \mathcal{L}\big(g \circ f(x, \theta_0 + \mathbf{B}\mathbf{A}),\, h(x)\big). \quad (5)$$

The expanded update $\tilde{\tau} = \mathbf{B}\mathbf{A}$ is the corresponding full-space parameter variation, while the concatenated matrix

$$\mathbf{B} = [\mathbf{B}_1, \ldots, \mathbf{B}_L] \in \mathbb{R}^{m\times(rL)} \quad (6)$$

is used as the compact PAVE specification. For identification, the learnware dock system can directly compute cosine similarity in the low-rank space via $\cos(\mathbf{B}, \mathbf{B}')$, which efficiently approximates $\cos(\mathbf{B}\mathbf{A}, \mathbf{B}'\mathbf{A})$ with high probability. In the compact-release regime analyzed in this paper, the public specification is the compact factor $\mathbf{B}$; the realized random factors $\mathbf{A}_\ell$ are not part of the public specification.

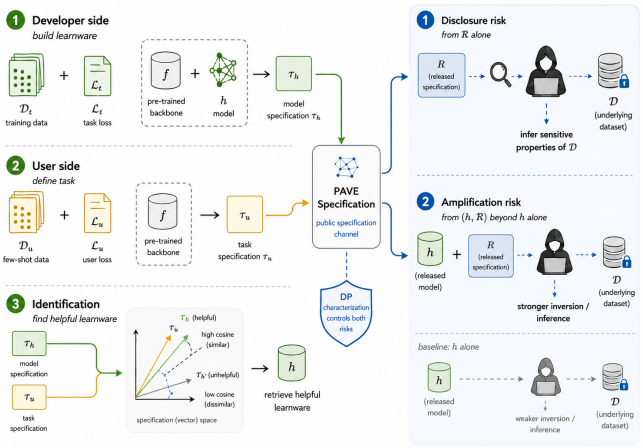

Figure 1: Two-stage learnware paradigm and privacy risks.

2.2 Formalization of Learnware Privacy Concerns

Since specifications are generated from the raw data of developers or users, a central privacy question is: *what additional risks are introduced by attaching and releasing a specification?* We distinguish two complementary concerns below.

(i) Disclosure risk. The released specification itself may leak information about the underlying dataset. This concern is particularly relevant for PAVE, because it summarizes the fine-tuning dataset through gradient-induced parameter variations. Prior studies on gradient and update inversion show that such information can be used to infer sensitive properties of training examples [Zhu *et al.*, 2019; Geiping *et al.*, 2020].

(ii) Amplification risk. Even without a specification, the released model may already leak information about its training data through model inversion attacks, which attempt to recover sensitive attributes or representative training examples from model outputs or parameters [Fredrikson *et al.*, 2014, 2015; Zhang *et al.*, 2020]. In a learnware system, publishing a specification alongside the model introduces an additional evidence channel. The learnware-specific risk we study is whether this extra channel can strengthen such attacks beyond what the model alone already enables.

We next formalize these two risks through game-based definitions. The goal is not to commit to a particular attack implementation, but to capture the adversary's view, side information, and inference objective in a common form.

Attack goals. We focus on two representative inference goals that are widely studied in the privacy literature and are natural in learnware systems:

- **Membership inference:** determine whether a particular sample appears in the developer's training set.
- **Attribute inference:** infer a sensitive attribute of a sample given its non-sensitive features.

A generic inference game. Let D denote the developer's training dataset, drawn from an underlying distribution $\mathcal{P}$. A learnware consists of a released model h together with a released specification R. We model an attack by a tuple $(\mathcal{A}, \tau, \nu)$, where $\mathcal{A}$ is a possibly randomized adversary, τ is a target function encoding the inference objective, and ν denotes the side information provided to the adversary. The interaction proceeds as follows:

(1) The challenger samples a record s from $\mathcal{P}$, or equivalently from D when modeling membership.
(2) The challenger reveals the side information $\nu(s)$ to the adversary.
(3) The adversary is given a released interface, specified below, and outputs a guess.

The adversary's success is defined as

$$\mathrm{gain}(\mathcal{A}; \mathcal{V}, \tau, \nu) \triangleq \Pr\left[\mathcal{A}(\mathcal{V}, \nu(s)) = \tau(s)\right], \tag{7}$$

where the probability is over the randomness of the sampled record, the released interface, and the adversary. Different privacy concerns correspond to different adversarial views $\mathcal{V}$.

(i) Disclosure risk of releasing the specification. We first quantify what the specification R reveals by itself. In this game, the adversary observes the released specification, denoted by $\mathcal{V}_{\mathrm{spec}} = R$. Since side information may already reveal part of the target, we define disclosure risk as the incremental advantage brought by releasing R:

$$\mathrm{Risk}_{\mathrm{dis}}(R) \triangleq \sup_{\mathcal{A}\in\mathfrak{A}} \mathrm{gain}(\mathcal{A}; R, \tau, \nu) - \sup_{\mathcal{A}\in\mathfrak{A}} \mathrm{gain}(\mathcal{A}; \varnothing, \tau, \nu), \tag{8}$$

where $\mathfrak{A}$ denotes the class of admissible adversaries, and $\varnothing$ denotes the baseline view in which the adversary receives only the side information. A small $\mathrm{Risk}_{\mathrm{dis}}(R)$ indicates that observing the released specification does not substantially improve the adversary's ability to infer $\tau(s)$ beyond what is already possible from side information.

(ii) Amplification risk for attacks against the released model. Releasing a specification may also help an attacker exploit the *model* more effectively. To capture this, we compare adversaries that observe the model alone with adversaries that observe both the model and the specification. Let $\mathcal{V}_{\mathrm{model}} = h$ denote access to the released model alone, and let $\mathcal{V}_{\mathrm{joint}} = (h, R)$ denote access to both. We define the *amplification risk* induced by R for the model h as

$$\begin{aligned}\mathrm{Risk}_{\mathrm{amp}}(h; R) \triangleq &\sup_{\mathcal{A}\in\mathfrak{A}} \mathrm{gain}(\mathcal{A}; (h, R), \tau, \nu) \\ &- \sup_{\mathcal{A}\in\mathfrak{A}} \mathrm{gain}(\mathcal{A}; h, \tau, \nu).\end{aligned} \tag{9}$$

This quantity isolates the incremental privacy risk introduced by publishing the specification on top of the released model.

Binary neighbouring-dataset instantiation. The gain-based definitions above describe the operational meaning of the two risks for general inference objectives. To obtain attack-agnostic theoretical guarantees, we use their standard binary neighbouring-dataset instantiation. In this instantiation, the adversary distinguishes whether a released view is generated from a dataset D or from a neighbouring dataset D'. This captures the worst-case form of membership- and attribute-style inference while avoiding commitment to a particular attack algorithm or target distribution.

For a released view V, define

$$p^V_{\mathcal{A}}(D) \triangleq \Pr\left[\mathcal{A}(V(D)) = 1\right], \tag{10}$$

and the optimal distinguishing advantage

$$\mathrm{Adv}^{\star}(V; D, D') \triangleq \sup_{\mathcal{A}\in\mathfrak{A}} \left| p^V_{\mathcal{A}}(D) - p^V_{\mathcal{A}}(D') \right|. \tag{11}$$

Equivalently, in the balanced binary game between D and D', the optimal success probability is

$$\mathrm{gain}^{\star}_{\mathrm{bin}}(V; D, D') = \frac{1}{2} + \frac{1}{2}\,\mathrm{Adv}^{\star}(V; D, D'). \tag{12}$$

Thus, controlling $\mathrm{Adv}^{\star}$ also controls the corresponding binary inference gain. Let $[x]_+ \triangleq \max\{x, 0\}$. The advantage-style disclosure risk $\mathrm{Risk}^{\mathrm{adv}}_{\mathrm{dis}}(R \mid H_0)$ is

$$\sup_{D\sim D'} \Big[\mathrm{Adv}^{\star}((H_0, R); D, D') - \mathrm{Adv}^{\star}(H_0; D, D')\Big]_+, \tag{13}$$

where H_0 denotes the baseline side information before observing the specification. Similarly, the advantage-style amplification risk $\mathrm{Risk}^{\mathrm{adv}}_{\mathrm{amp}}(H; R)$ is

$$\sup_{D\sim D'} \Big[\mathrm{Adv}^{\star}((H, R); D, D') - \mathrm{Adv}^{\star}(H; D, D')\Big]_+, \tag{14}$$

where H denotes the model-only view. These quantities are the binary neighbouring-dataset versions of the two game-based risks above, and are the objects bounded in our theoretical analysis.

3 DP Guarantees for PAVE Specifications

In this section, we characterize when PAVE specifications admit differential privacy. We first show that, under natural compact-release conditions of learnware docks, the compact PAVE factor satisfies an intrinsic (ε, δ)-DP guarantee without injecting additional additive noise. We then discuss the structural conditions behind this guarantee and introduce DP-S-PAVE, a certified variant for deployments outside the intrinsic compact-release regime.

We first recall some basic notions of differential privacy (DP). DP limits how much the output distribution can change when a member of the original dataset is modified.

Definition 3.1 (Neighbouring datasets)**.** *Let $D = \{z_i\}_{i=1}^N$ and $D' = \{z_i'\}_{i=1}^N$ be two datasets of the same size. We write $D \sim D'$ if they differ in exactly one record, i.e., there exists an index j such that $z_j \neq z_j'$ and $z_i = z_i'$ for all $i \neq j$.*

Definition 3.2 (Differential privacy)**.** *A randomized mechanism $\mathcal{M}$ is (ε, δ)-differentially private if for any measurable event E and any neighbouring datasets $D \sim D'$,*

$$\Pr\big(\mathcal{M}(D) \in E\big) \leq e^{\varepsilon} \Pr\big(\mathcal{M}(D') \in E\big) + \delta, \tag{15}$$

where the probability is taken over the randomness of $\mathcal{M}$.

Definition 3.3 (ℓ_2-sensitivity)**.** *Let q be a matrix-valued query. Its ℓ_2-sensitivity is*

$$\Delta(q) \triangleq \max_{D \sim D'} \big\| q(D) - q(D') \big\|_2, \tag{16}$$

where $\|\cdot\|_2$ denotes the spectral norm.

From PAVE to a compact Gaussian-sketch mechanism. For clarity, we first consider a single layer and omit the layer index. Let $\mathbf{A} \in \mathbb{R}^{r \times m}$ be sampled with i.i.d. entries $A_{ij} \sim \mathcal{N}(0, 1/r)$, independently of the dataset. In the compact privacy regime, the distribution of $\mathbf{A}$ is public, while its realized value and random seed are maintained as internal randomness and are not part of the public specification.

For one-step, linearized, or trajectory-decoupled PAVE, the learned compact factor can be written as

$$\mathbf{B}(D) = \mathbf{F}(D)\mathbf{A}^\top, \tag{17}$$

where $\mathbf{F}(D) \in \mathbb{R}^{d \times m}$ is the matrix query induced by the fine-tuning trajectory. For example, unrolling gradient steps of the form $\mathbf{B}_{t+1} = \mathbf{B}_t - \eta\, \mathbf{G}_t(D)\mathbf{A}^\top$ gives $\mathbf{B}_T(D) = -\eta \sum_{t=0}^{T-1} \mathbf{G}_t(D)\mathbf{A}^\top$, so that $\mathbf{F}(D) = -\eta \sum_{t=0}^{T-1} \mathbf{G}_t(D)$ when the cumulative query is independent of the realized $\mathbf{A}$.

The mechanism in Eq. (17) is a Gaussian sketch. To avoid support separation for matrix-valued releases, we analyze the standard fixed-subspace regime. Assume that there exists a fixed and public matrix $\mathbf{U} \in \mathbb{R}^{d \times k}$, $\mathbf{U}^\top \mathbf{U} = \mathbf{I}_k$, such that the released query lies in this specification subspace:

$$\mathbf{F}(D) = \mathbf{U}\mathbf{H}(D), \qquad \mathbf{H}(D) \in \mathbb{R}^{k \times m}. \tag{18}$$

Equivalently, the released compact factor can be written as

$$\mathbf{B}(D) = \mathbf{U}\mathbf{C}(D), \qquad \mathbf{C}(D) = \mathbf{H}(D)\mathbf{A}^\top. \tag{19}$$

Since $\mathbf{U}$ is fixed and public, it suffices to analyze the coordinate release $\mathbf{C}(D)$. We define the covariance matrix

$$\mathbf{\Sigma}_D \triangleq \mathbf{H}(D)\mathbf{H}(D)^\top \in \mathbb{R}^{k \times k}. \tag{20}$$

We impose the following covariance-stability condition on admissible neighbouring datasets:

$$(1-\gamma)\mathbf{\Sigma}_{D'} \preceq \mathbf{\Sigma}_D \preceq (1+\gamma)\mathbf{\Sigma}_{D'}, \qquad 0 < \gamma < 1, \tag{21}$$

where $\mathbf{\Sigma}_D$ and $\mathbf{\Sigma}_{D'}$ are positive definite. This condition states that changing one record cannot substantially rotate or rescale the covariance inside the specification subspace.

Definition 3.4 (Rényi differential privacy)**.** *For a concise privacy accounting, we use Rényi differential privacy (RDP). For $\alpha > 1$, a mechanism $\mathcal{M}$ is said to satisfy (α, ρ)-RDP if*

$$D_\alpha(\mathcal{M}(D) \| \mathcal{M}(D')) \leq \rho \tag{22}$$

for all neighbouring datasets $D \sim D'$, where $D_\alpha(\cdot\|\cdot)$ is the order-α Rényi divergence. An (α, ρ)-RDP mechanism is also $\Big(\rho + \frac{\log(1/\delta)}{\alpha - 1},\, \delta\Big)$-DP for any $\delta \in (0, 1)$. In particular, a $(2, \rho_2)$-RDP guarantee gives $(\rho_2 + \log(1/\delta), \delta)$-DP.

Theorem 3.5 (Intrinsic DP for compact PAVE specifications)**.** *Consider the compact PAVE mechanism*

$$\mathcal{M}_{\mathbf{B}}(D) = \mathbf{B}(D) = \mathbf{U}\mathbf{H}(D)\mathbf{A}^\top, \tag{23}$$

where $A_{ij} \sim \mathcal{N}(0, 1/r)$ is sampled independently of D, and the realized $\mathbf{A}$ is not released. Assume the common-support condition in Eq. (18) and the covariance-stability condition in Eq. (21). Then $\mathcal{M}_{\mathbf{B}}$ satisfies $(2, \rho_2)$-RDP with

$$\rho_2 \leq \frac{rk}{2} \log \frac{1}{1-\gamma^2} \leq \frac{rk\gamma^2}{2(1-\gamma^2)}. \tag{24}$$

Consequently, for any $\delta \in (0, 1)$, $\mathcal{M}_{\mathbf{B}}$ is (ε, δ)-DP with

$$\varepsilon = \frac{rk}{2} \log \frac{1}{1-\gamma^2} + \log \frac{1}{\delta} \leq \frac{rk\gamma^2}{2(1-\gamma^2)} + \log \frac{1}{\delta}. \tag{25}$$

Remark 3.6. *Theorem 3.5 characterizes an intrinsic DP regime of compact PAVE specifications. It relies on three structural conditions and one modeling scope condition.*

Compact secret release. *The public PAVE specification is the compact factor $\mathbf{B} = \mathbf{F}(D)\mathbf{A}^\top$, while the realized Gaussian matrix $\mathbf{A}$, its seed, and the expanded update $\mathbf{BA}$ are not part of the public release. The distribution of $\mathbf{A}$ can be public; what is not released is the particular realization used to generate the specification. This distinction is essential because $\mathbf{A}$ serves as the internal randomness of the compact Gaussian-sketch mechanism. If $(\mathbf{A}, \mathbf{B})$ is released, the algebraic relation $\mathbf{B} = \mathbf{F}(D)\mathbf{A}^\top$ becomes directly testable, and the Gaussian-sketch privacy argument no longer applies. The modeling scope condition is related but different: the matrix query $\mathbf{F}(D)$ should be independent of the realized $\mathbf{A}$. This holds for one-step or trajectory-decoupled PAVE, where the cumulative query is fixed once D is fixed. For fully adaptive PAVE training, the certified variant below should be used.*

Common specification subspace. *The second structural condition is the fixed public subspace in Eq. (18). Its role is to place all released PAVE specifications in the same public coordinate system: $\mathbf{F}(D) = \mathbf{U}\mathbf{H}(D)$ and $\mathbf{B}(D) = \mathbf{U}(\mathbf{H}(D)\mathbf{A}^\top)$. This condition is natural in a learnware dock, because model-side and user-side specifications must be comparable for identification. It is automatically satisfied when*

the selected PAVE query has stable full row support, in which case one may take $\mathbf{U} = \mathbf{I}$*. More generally, the system can enforce it by selecting fixed layers, using a fixed public projection subspace, choosing a subspace from public calibration data, or directly releasing the coordinate specification* $\mathbf{C}(D) = \mathbf{U}^\top \mathbf{B}(D) = \mathbf{H}(D)\mathbf{A}^\top$*. Without such a common support, neighbouring datasets may move the released matrix into different column supports, causing support separation.*

Covariance stability. *The third structural condition is covariance stability in Eq.* (21)*. It requires the Gaussian-sketch covariance* $\mathbf{\Sigma}_D = \mathbf{H}(D)\mathbf{H}(D)^\top$ *to change smoothly under a single-record modification. The positive definiteness of* $\mathbf{\Sigma}_D$ *is the non-degeneracy part of this condition: the retained public specification subspace should not collapse to a lower-dimensional direction. A convenient sufficient condition is* $\mu^2 \mathbf{I}_k \preceq \mathbf{H}(D)\mathbf{H}(D)^\top \preceq L^2 \mathbf{I}_k$ *and* $\|\mathbf{H}(D) - \mathbf{H}(D')\|_2 \le \Delta$ *for all neighbouring* $D \sim D'$*. Then* $\|\mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'}\|_2 \le 2L\Delta$*, and Eq.* (21) *holds with*

$$\gamma = \frac{2L\Delta}{\mu^2}, \tag{26}$$

provided $\gamma < 1$*. For an average-type query* $\mathbf{H}(D) = n^{-1}\sum_{i=1}^n \mathbf{\Psi}(z_i)$ *with* $\|\mathbf{\Psi}(z)\|_2 \le G$*, one has* $\Delta \le 2G/n$*, and hence* $\gamma \le 4LG/(n\mu^2)$*. Thus large fine-tuning datasets, stable optimization, and clipping of unusually large updates all make the intrinsic compact guarantee more favorable.*

Proposition 3.7 (A public-$\mathbf{A}$ compact release is not generally DP)**.** *Consider the mechanism*

$$\mathcal{M}_{\mathrm{pub}}(D) = (\mathbf{A},\, \mathbf{F}(D)\mathbf{A}^\top), \tag{27}$$

where $\mathbf{A} \in \mathbb{R}^{r\times m}$ *has i.i.d. Gaussian entries and the realized* $\mathbf{A}$ *is released. If there exist neighbouring datasets* $D \sim D'$ *such that* $\mathbf{F}(D) - \mathbf{F}(D') \neq \mathbf{0}$*, then, for such a neighbouring pair,* $\mathcal{M}_{\mathrm{pub}}$ *is not* (ε, δ)*-DP for any finite* ε *and any* $\delta < 1$*.*

Proposition 3.7 explains why Theorem 3.5 treats the realized $\mathbf{A}$ as internal randomness. The proposition does not say that the sampling law of $\mathbf{A}$ must be hidden; the distribution, dimension, and generation procedure of $\mathbf{A}$ can be public. What must remain internal in the intrinsic compact-release regime is the actual realization, or equivalently the seed that reproduces it. If a deployment needs to publish $\mathbf{A}$, its seed, or the expanded update $\mathbf{BA}$, then the certified DP-S-PAVE variant below should be used instead.

Remark 3.8. *The quantities in Theorem 3.5 have direct learnware interpretations. The rank* r *is the low-rank width of the PAVE factor, and* k *is the dimension of the public specification subspace. Their product* rk *is the effective number of Gaussian sketch coordinates released to the system. The parameter* γ *measures the relative change of the specification covariance under a neighbouring-dataset modification. Thus, smaller effective specification dimension and more stable task-level parameter variations lead to tighter privacy. These quantities also correspond to design choices of the system: the rank controls how much parameter-variation signal is retained, the public subspace controls the common coordinate system for matching, and the stability parameter reflects the influence of individual records on the specification.*

Algorithm 1 DP-Stabilized PAVE (DP-S-PAVE)

Input: dataset D; pretrained parameters θ_0; selected layers $\mathcal{L}$; ranks $\{r_\ell\}$; steps T; step size η; clipping threshold C; sampling rate q or batch size b; noise multiplier σ.
Output: private compact specification $R = \{\mathbf{B}_{\ell,T}\}_{\ell\in\mathcal{L}}$.
for $\ell \in \mathcal{L}$ **do**
  Initialize $\mathbf{A}_\ell \in \mathbb{R}^{r_\ell \times n_\ell}$ independently of D and freeze it; set $\mathbf{B}_{\ell,0} \leftarrow \mathbf{0}$.
end for
for $t = 0, \dots, T-1$ **do**
  Sample a mini-batch $S_t \subseteq D$ with rate q or $|S_t| \approx b$.
  for $\ell \in \mathcal{L}$ **do**
    for $z_i \in S_t$ **do**
      $\mathbf{Z}_{\ell,t}^{(i)} \leftarrow \nabla_{\mathbf{W}_\ell} \mathcal{L}(\mathbf{W}_t; z_i)\mathbf{A}_\ell^\top$; $\widetilde{\mathbf{Z}}_{\ell,t}^{(i)} \leftarrow \mathrm{Clip}(\mathbf{Z}_{\ell,t}^{(i)}, C)$.
    end for
    Draw $\mathbf{N}_{\ell,t}$ with i.i.d. entries from $\mathcal{N}(0, \sigma^2 C^2)$.
    $\bar{\mathbf{Z}}_{\ell,t} \leftarrow b^{-1}\Big(\sum_{z_i\in S_t} \widetilde{\mathbf{Z}}_{\ell,t}^{(i)} + \mathbf{N}_{\ell,t}\Big)$.
    $\mathbf{B}_{\ell,t+1} \leftarrow \mathbf{B}_{\ell,t} - \eta \bar{\mathbf{Z}}_{\ell,t}$.
  end for
end for
return $R = \{\mathbf{B}_{\ell,T}\}_{\ell\in\mathcal{L}}$.

A certified variant beyond the intrinsic regime. The above structural conditions are natural in a compact-release regime, but they need not hold in every deployment. For example, a system may require public reproducibility of $\mathbf{A}$, may need to release the expanded update $\mathbf{BA}$, may not enforce a common specification subspace, or may use fully adaptive nonlinear PAVE training in which the cumulative query depends on the realized $\mathbf{A}$. To cover such cases, we introduce a certified differentially private variant, called DP-Stabilized PAVE (DP-S-PAVE). The idea is to apply per-example clipping (Define $\mathrm{Clip}(\mathbf{Z}, C) = \mathbf{Z}\cdot\min\{1, C/\|\mathbf{Z}\|_{\mathrm{F}}\}$) and calibrated Gaussian noise directly in the low-rank PAVE update space. In contrast to Theorem 3.5, DP-S-PAVE does not rely on secret $\mathbf{A}$, common support, or covariance stability; its guarantee follows from standard DP-SGD/RDP accounting.

Theorem 3.9 (Certified DP of DP-S-PAVE)**.** *Fix the matrices* $\{\mathbf{A}_\ell\}_{\ell\in\mathcal{L}}$ *independently of the dataset. In Algorithm 1, assume that every per-example projected update is clipped to Frobenius norm at most* C*, and that independent Gaussian noise with standard deviation* σC *is added before the deterministic scaling by* $1/b$*. Let* $\rho_\alpha^{\mathrm{subG}}(q,\sigma)$ *denote the order-*α *RDP parameter of one Poisson-subsampled Gaussian update with replacement-neighbour sensitivity* $2C$ *and noise standard deviation* σC*. Then the final release* $R = \{\mathbf{B}_{\ell,T}\}_{\ell\in\mathcal{L}}$ *satisfies* (α, ρ_α)*-RDP with*

$$\rho_\alpha \le \sum_{t=0}^{T-1}\sum_{\ell\in\mathcal{L}} \rho_\alpha^{\mathrm{subG}}(q,\sigma). \tag{28}$$

Consequently, for any $\delta \in (0,1)$*, DP-S-PAVE is* $(\varepsilon_{\mathrm{dp}}, \delta)$*-DP with*

$$\varepsilon_{\mathrm{dp}}(\delta) = \inf_{\alpha>1}\left\{\rho_\alpha + \frac{\log(1/\delta)}{\alpha-1}\right\}. \tag{29}$$

In particular, without subsampling, i.e., $q = 1$*, a conservative closed-form bound is*

$$\rho_\alpha \le \frac{2\alpha T|\mathcal{L}|}{\sigma^2}, \qquad \varepsilon_{\mathrm{dp}} \le \frac{2T|\mathcal{L}|}{\sigma^2} + 2\sqrt{\frac{2T|\mathcal{L}|}{\sigma^2}\log\frac{1}{\delta}}. \tag{30}$$

Remark 3.10. *DP-S-PAVE complements the intrinsic compact guarantee. When the compact secret-sketch conditions hold, Theorem 3.5 gives a noise-free DP route. When these conditions are not enforced, Theorem 3.9 provides a certified route by adding calibrated noise in the low-rank PAVE update space. Since the guarantee of DP-S-PAVE is conditioned on fixed $\mathbf{A}_\ell$ and comes from explicit Gaussian perturbation, the realized $\mathbf{A}_\ell$, the compact factors $\mathbf{B}_\ell$, the expanded updates $\mathbf{B}_\ell\mathbf{A}_\ell$, normalized specifications, cosine similarities, and top-K retrieval outputs are all post-processing of a DP mechanism and therefore remain differentially private.*

4 Risk Analysis

We now connect the DP guarantees in Section 3 to the two learnware privacy risks formalized in Section 2.2, and then discuss the resulting privacy–utility trade-off. The key message is that the released specification channel is the object whose privacy parameters control both risks.

4.1 Disclosure Risk of the Specification Channel

We first consider the disclosure risk of releasing the specification itself. For brevity, we write

$$\eta_{\varepsilon,\delta} \triangleq (e^{\varepsilon} - 1) + \delta. \tag{31}$$

Lemma 4.1 (DP channels add bounded inference advantage)**.** *Let H be a baseline view, and let $R_D = \mathcal{M}(D)$ be an additional released channel. Suppose that R is conditionally (ε, δ)-DP given H, i.e., for every fixed value $H = u$, every measurable event S, and all $D \sim D'$,*

$$\Pr\left[R_D \in S \mid H = u\right] \le e^{\varepsilon} \Pr\left[R_{D'} \in S \mid H = u\right] + \delta. \tag{32}$$

Then, for all neighbouring datasets $D \sim D'$,

$$\mathrm{Adv}^{\star}((H, R); D, D') \le \mathrm{Adv}^{\star}(H; D, D') + \eta_{\varepsilon,\delta}. \tag{33}$$

Lemma 4.1 is the link between mechanism-level privacy and learnware-level risk. After the view $H = u$ is fixed, any attack using (u, R) is a post-processing of the released specification. The term $\mathrm{Adv}^{\star}(H; D, D')$ accounts for the distinguishing power already present in the baseline view, while the additional contribution of R is controlled by $\eta_{\varepsilon,\delta}$.

For disclosure risk, the view H_0 contains the side information available before observing the specification, and the additional channel is the released PAVE specification R. Thus the disclosure question is whether observing R gives substantially more distinguishing power than the baseline view alone.

Theorem 4.2 (DP controls specification disclosure risk)**.** *Let $R = \mathcal{M}_{\mathbf{B}}(D)$ be the released compact PAVE specification. If R is conditionally (ε, δ)-DP given the baseline side information H_0, then*

$$\mathrm{Risk}^{\mathrm{adv}}_{\mathrm{dis}}(R \mid H_0) \le (e^{\varepsilon} - 1) + \delta. \tag{34}$$

In particular, the same bound holds when H_0 is fixed public side information and R satisfies ordinary (ε, δ)-DP.

Theorem 4.2 is an incremental-risk statement. It does not bound the raw success probability of an attack, because the baseline side information may already reveal part of the target. Instead, it bounds the extra binary distinguishing advantage attributable to observing the specification. Consequently, tighter compact-release privacy parameters, smaller effective dimension rk, stronger covariance stability, or the use of DP-S-PAVE directly reduce the worst-case disclosure advantage.

4.2 Specification-Side Amplification Risk

We next consider amplification risk: whether publishing a PAVE specification can strengthen attacks against the released model. Let H denote the model-only view available to the adversary, including the released model h and any side information shared across the two games. The joint view is (H, R). The model view H may already contain information about the underlying dataset; our goal is to isolate the additional contribution of the specification channel.

We assume that the compact specification channel R is conditionally (ε, δ)-DP given H. This is satisfied, for example, when H is treated as fixed public side information and the compact PAVE release satisfies Theorem 3.5 uniformly under this conditioning. If the model h is trained by a non-private procedure, it may still leak information, but the result below controls the incremental leakage caused by adding R.

Theorem 4.3 (No material specification-side amplification)**.** *Let $R = \mathcal{M}_{\mathbf{B}}(D)$ be the released compact PAVE specification, and let H be the model-only view. If R is conditionally (ε, δ)-DP given H, then*

$$\mathrm{Risk}^{\mathrm{adv}}_{\mathrm{amp}}(H; R) \le (e^{\varepsilon} - 1) + \delta. \tag{35}$$

Although Theorem 4.3 has the same numerical margin as the disclosure bound, it controls a different learnware-specific risk. The disclosure bound compares the baseline side information H_0 with the joint view (H_0, R), and therefore measures what the specification reveals by itself. In contrast, the amplification bound takes the model-only view H as the baseline. This view may already be data-dependent and non-private, and the theorem does not attempt to remove such model-side leakage. Instead, it asks whether the additional specification channel R can further increase the adversary's distinguishing power beyond what is already possible from H. Technically, both results rely on Lemma 4.1, but they instantiate it with different baselines: H_0 for disclosure and H for amplification. The identical margin $(e^{\varepsilon} - 1) + \delta$ reflects that, in both cases, the only newly added channel is the DP-protected specification. Thus Theorem 4.3 should be read as a specification-side amplification guarantee: publishing PAVE may be combined with model access, but the extra inference advantage attributable to the specification is bounded by the DP-controlled margin in Eq. (35).

This is a specification-side guarantee. It separates the information already present in the released model view from the additional risk introduced by the specification. If a deployment releases objects outside the compact secret-release regime, such as the realized $\mathbf{A}$, the expanded update $\mathbf{BA}$, or an adapted predictor constructed from a public $\mathbf{A}$, then Theorem 3.5 may no longer apply. In such cases, the same disclosure and amplification conclusions can be obtained by using the certified DP-S-PAVE guarantee in Theorem 3.9. Once the specification channel is DP, downstream dock operations such as normalization, cosine similarity, ranking, and top-K retrieval remain protected by post-processing.

4.3 Privacy–Utility Trade-off of PAVE

The preceding results show that both disclosure risk and specification-side amplification risk are controlled by the DP parameters of the released specification channel. Theorem 3.5 makes these parameters explicit for compact PAVE. In the intrinsic regime, the order-2 RDP parameter satisfies

$$\rho_2 \leq \frac{rk}{2} \log \frac{1}{1-\gamma^2}. \tag{36}$$

Thus the effective released dimension rk and the covariance-stability parameter γ are the main privacy-relevant quantities. The same quantities also affect learnware identification. The rank r controls the width of the low-rank PAVE factor, while k controls the dimension of the public specification subspace. Larger r or k can preserve richer parameter-variation signals and improve matching, but they also release more Gaussian sketch coordinates and increase the privacy cost. Conversely, overly small r or k may improve privacy but remove directions needed to distinguish helpful learnwares.

The stability parameter γ captures another side of the trade-off. It measures how much the covariance $\mathbf{\Sigma}_D = \mathbf{H}(D)\mathbf{H}(D)^\top$ of the projected PAVE query changes when one record is modified. When $\mathbf{H}(D)$ is stable and non-degenerate, γ is small and the privacy bound becomes tighter. The sufficient condition in Section 3 gives a concrete interpretation: if $\mu^2\mathbf{I}_k \preceq \mathbf{H}(D)\mathbf{H}(D)^\top \preceq L^2\mathbf{I}_k$ and $\|\mathbf{H}(D) - \mathbf{H}(D')\|_2 \leq \Delta$, then $\gamma = 2L\Delta/\mu^2$; for average-type queries with bounded per-example contribution, $\Delta = O(1/n)$. Larger fine-tuning datasets, stable optimization, and clipping of unusually large updates therefore reduce the effect of any individual record on the specification.

This stability is also useful for identification. Learnware matching should depend on task-level variation rather than example-specific spikes. Suppressing unstable record-level influence helps the PAVE specification reflect reusable capability information instead of idiosyncratic noise, which can improve the robustness of cosine-based comparisons across learnwares. In this sense, privacy and utility are not always opposed: the same stabilization that reduces γ can also make the specification more faithful to the task-level signal.

The public specification subspace also has a dual role. From the privacy perspective, a fixed subspace $\mathbf{U}$ prevents support separation across neighbouring datasets and enables the compact Gaussian-sketch analysis. From the learnware perspective, $\mathbf{U}$ provides a common coordinate system in which model-side and user-side specifications can be compared. In practice, $\mathbf{U}$ may be chosen by fixed layer selection or a public projection, while the effective dimension k can be tuned according to the desired privacy–utility balance.

When the compact secret-release conditions are not suitable, DP-S-PAVE provides the certified alternative. Its privacy is controlled by the clipping threshold, noise multiplier, sampling rate, and number of optimization steps through standard DP accounting. Since the perturbation is applied in the low-rank PAVE update space rather than in the full parameter space, the certified variant still respects the lightweight nature of learnware specifications while providing formal protection beyond the intrinsic compact-release regime.

5 Related Work

Most existing learnware systems build with RKME as a specification choice [Zhou and Tan, 2024; Wu *et al.*, 2023]. This has led to a rich line of work on learnware identification, heterogeneous feature or label spaces, and evolvable learnware systems [Tan *et al.*, 2026; Liu *et al.*, 2024; Tan *et al.*, 2024]. On the privacy side, recent work provides rigorous guarantees for RKME-style specifications through geometric analysis of reduced sets [Lei *et al.*, 2024]. Our work studies the same general objective of privacy of learnware specifications, but for a fundamentally different specification form. PAVE summarizes model capability through parameter variations induced by fine-tuning, rather than through representatives in the data space. This difference changes both the privacy surface and the mathematical tools needed for analysis.

Our analysis is also related to DP mechanisms for learning and sketching. Standard learning-based DP guarantees are usually obtained by explicit noise injection, such as DP-SGD, RDP, and subsampled-RDP accounting [Abadi *et al.*, 2016; Mironov, 2017; Wang *et al.*, 2019]. Another line shows that random projections or Gaussian sketches can themselves provide privacy guarantees [Blocki *et al.*, 2012; Lev *et al.*, 2025]. Recent work on LoRA privacy studies Wishart projection mechanisms and shows that matrix-valued LoRA-style updates are not generally noise-free DP [Hu *et al.*, 2026]. Our compact PAVE result is different: under dock-side compact-release conditions, the released factor is analyzed as a Gaussian sketch of stable parameter variations, and DP-S-PAVE recovers certified protection by explicit perturbation.

Finally, our risk formulation is motivated by privacy attacks on learning systems. Gradient and update inversion motivate the disclosure risk of releasing PAVE [Zhu *et al.*, 2019; Geiping *et al.*, 2020], while membership, attribute, and model inversion attacks motivate the amplification question when a specification is released together with a model [Fredrikson *et al.*, 2014, 2015; Shokri *et al.*, 2017; Yeom *et al.*, 2018; Salem *et al.*, 2019]. Our DP-to-advantage analysis connects these risks to the privacy of the released specification.

6 Concluding Remarks

This paper provides the first formal privacy characterization of PAVE specifications in the learnware paradigm. We show that, under natural compact-release conditions of learnware docks, compact PAVE specifications admit an explicit (ε, δ)-differential privacy guarantee through a Gaussian-sketch view of stable parameter variations, without introducing explicit additive noise. For deployments outside this intrinsic regime, we further provide DP-S-PAVE as a certified differentially private variant. We connect these DP guarantees to rigorous bounds on both disclosure risk and specification-side amplification risk, showing that the additional inference advantage attributable to the specification is controlled by the corresponding DP parameters. Finally, we discuss how privacy and identification utility vary with the PAVE generation settings. Our analysis offers useful building blocks for privacy-aware learnware specification design and contributes to a DP-based understanding of dataset leakage through parameter-variation representations.

Ethical Statement

This work studies privacy risks and privacy-preserving guarantees for PAVE specifications in learnware systems. It does not collect new human-subject data or conduct experiments on private user data. The proposed analysis is intended to support privacy-aware deployment by clarifying when compact PAVE releases admit formal differential privacy guarantees and by providing DP-S-PAVE for settings where the structural conditions are not enforced. We also emphasize that practical deployments should verify the required assumptions or use the certified DP variant before releasing specifications.

Acknowledgments

This work was supported by FIDBP of MoE (JYB2025XDXM118) and 111 Center (B26023).

Tan was supported by the Postdoctoral Fellowship Program of CPSF (GZB20250396) and Jiangsu Funding Program for Excellent Postdoctoral Talent.

References

Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The johnson-lindenstrauss transform itself preserves differential privacy. In *Proceedings of the 53rd Annual IEEE Symposium on Foundations of Computer Science*, pages 410–419, 2012.

Cynthia Dwork. Differential privacy. In *Proceedings of International Colloquium on Automata, Languages, and Programming*, pages 1–12, 2006.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: an end-to-end case study of personalized warfarin dosing. In *Proceedings of the 23rd USENIX Conference on Security Symposium*, pages 17–32, 2014.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333, 2015.

Jonas Geiping, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. Inverting gradients: How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems 33*, pages 16937–16947, 2020.

Yaxi Hu, Johanna Düngler, Bernhard Schölkopf, and Amartya Sanyal. LoRA and Privacy: When Random Projections Help (and When They Don't), 2026. arXiv preprint arXiv:2601.21719.

Hao-Yi Lei, Zhi-Hao Tan, and Zhi-Hua Zhou. On the ability of developers' training data preservation of learnware. In *Advances in Neural Information Processing Systems 37*, pages 36471–36513, 2024.

Omri Lev, Vishwak Srinivasan, Moshe Shenfeld, Katrina Ligett, Ayush Sekhari, and Ashia C. Wilson. The Gaussian Mixing Mechanism: Rényi Differential Privacy via Gaussian Sketches, 2025. arXiv preprint arXiv:2505.24603.

Jian-Dong Liu, Zhi-Hao Tan, and Zhi-Hua Zhou. Towards making learnware specification and market evolvable. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pages 13909–13917, 2024.

Ilya Mironov. Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium*, pages 263–275, 2017.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *Proceedings of the 26th Annual Network and Distributed System Security Symposium*, 2019.

Hao-Yu Shi, Zhi-Hao Tan, Zi-Chen Zhao, Yang Yu, and Zhi-Hua Zhou. A study on pave specification for learnware. In *The 14th International Conference on Learning Representations*, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of 38th IEEE Symposium on Security and Privacy*, pages 3–18, 2017.

Peng Tan, Zhi-Hao Tan, Yuan Jiang, and Zhi-Hua Zhou. Towards enabling learnware to handle heterogeneous feature spaces. *Machine Learning*, 113(4):1839–1860, 2024.

Peng Tan, Feifan Yang, Zhi-Hao Tan, and Zhi-Hua Zhou. Tabular learnwares can be repurposed for seemingly irrelevant new tasks. In *Proceedings of the 40th AAAI Conference on Artificial Intelligence*, number 30, pages 25778–25786, 2026.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pages 1226–1235, 2019.

Xi-Zhu Wu, Wenkai Xu, Song Liu, and Zhi-Hua Zhou. Model reuse with reduced kernel mean embedding specification. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):699–710, 2023.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *Proceedings of IEEE 31st Computer Security Foundations Symposium*, pages 268–282, 2018.

Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Zhi-Hua Zhou and Zhi-Hao Tan. Learnware: Small models do big. *Science China Information Sciences*, 67(1):112102, 2024.

Zhi-Hua Zhou. Learnware: on the future of machine learning. *Frontiers of Computer Science*, 10(4):589–590, 2016.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *Advances in Neural Information Processing Systems 32*, pages 14774–14784, 2019.

7 Proofs

This appendix provides complete proofs for the technical statements used in the main text. We first collect auxiliary facts about Rényi differential privacy, Gaussian mechanisms, post-processing, and composition. We then prove the compact Gaussian-sketch DP guarantee for PAVE specifications, the necessity of keeping the realized Gaussian factor internal, the certified DP guarantee of DP-S-PAVE, and finally the DP-to-risk implications for disclosure and amplification risks.

7.1 Notation

All vectors are viewed as column vectors. For a matrix $\mathbf{X}$, $\|\mathbf{X}\|_2$ denotes the spectral norm and $\|\mathbf{X}\|_{\mathrm{F}}$ denotes the Frobenius norm. For a symmetric matrix $\mathbf{S}$, $\lambda_i(\mathbf{S})$ denotes its i-th largest eigenvalue. For a general matrix $\mathbf{X}$, $\sigma_i(\mathbf{X})$ denotes its i-th largest singular value. For symmetric matrices, $\mathbf{X} \Pr eceq \mathbf{Y}$ means that $\mathbf{Y} - \mathbf{X}$ is positive semidefinite. All neighbouring-dataset assumptions are understood to hold for every ordered neighbouring pair $D \sim D'$.

7.2 Auxiliary Facts on RDP and DP

We first recall the basic RDP notation used in the paper. For two probability distributions P, Q with $P \ll Q$, the order-α Rényi divergence is

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\int\left(\frac{dP}{dQ}\right)^\alpha dQ, \qquad \alpha > 1. \quad (37)$$

A randomized mechanism $\mathcal{M}$ satisfies (α, ρ)-RDP if

$$D_\alpha(\mathcal{M}(D)\|\mathcal{M}(D')) \le \rho \quad (38)$$

for all neighbouring datasets $D \sim D'$.

Lemma 7.1 (RDP to approximate DP)**.** *If a mechanism $\mathcal{M}$ satisfies (α, ρ)-RDP for some $\alpha > 1$, then for any $\delta \in (0, 1)$, $\mathcal{M}$ is*

$$\left(\rho + \frac{\log(1/\delta)}{\alpha-1}, \delta\right)\text{-DP}. \quad (39)$$

In particular, if $\mathcal{M}$ satisfies $(2, \rho_2)$-RDP, then it is

$$\left(\rho_2 + \log\frac{1}{\delta}, \delta\right)\text{-DP}. \quad (40)$$

Proof. Fix neighbouring datasets $D \sim D'$ and write

$$P = \mathcal{L}(\mathcal{M}(D)), \qquad Q = \mathcal{L}(\mathcal{M}(D')). \quad (41)$$

Let

$$L(x) = \log\frac{dP}{dQ}(x) \quad (42)$$

be the privacy-loss random variable under P. By the definition of Rényi divergence,

$$\begin{aligned}&\mathbb{E}_P[\exp((\alpha-1)L)] = \\ &\exp((\alpha-1)D_\alpha(P\|Q)) \le \exp((\alpha-1)\rho)\,.\end{aligned} \quad (43)$$

For any measurable event E and any threshold $\varepsilon > 0$,

$$\begin{aligned}P(E) &= P(E \cap \{L \le \varepsilon\}) + P(E \cap \{L > \varepsilon\}) \\ &\le e^\varepsilon Q(E) + P(L > \varepsilon).\end{aligned} \quad (44)$$

By Markov's inequality,

$$\begin{aligned}P(L > \varepsilon) &= P(\exp((\alpha-1)L) > \exp((\alpha-1)\varepsilon)) \\ &\le \exp((\alpha-1)(\rho-\varepsilon)).\end{aligned} \quad (45)$$

Taking

$$\varepsilon = \rho + \frac{\log(1/\delta)}{\alpha-1} \quad (46)$$

gives

$$P(L > \varepsilon) \le \delta. \quad (47)$$

Therefore,

$$P(E) \le e^\varepsilon Q(E) + \delta. \quad (48)$$

Since this holds for all neighbouring datasets and measurable events, the mechanism is (ε, δ)-DP. □

Lemma 7.2 (Post-processing)**.** *Let $\mathcal{M}$ be an (ε, δ)-DP mechanism. Let ϕ be any possibly randomized mapping whose randomness is independent of the input dataset. Then $\phi \circ \mathcal{M}$ is also (ε, δ)-DP. The same statement holds for RDP with the same order and parameter.*

Proof. We prove the approximate-DP statement. The RDP statement follows from the data-processing inequality for Rényi divergence.

First suppose ϕ is deterministic. For any measurable event E in the output space of $\phi \circ \mathcal{M}$, define

$$S = \phi^{-1}(E). \quad (49)$$

Then

$$\Pr[(\phi \circ \mathcal{M})(D) \in E] = \Pr[\mathcal{M}(D) \in S]. \quad (50)$$

Since $\mathcal{M}$ is (ε, δ)-DP,

$$\Pr[\mathcal{M}(D) \in S] \le e^\varepsilon \Pr[\mathcal{M}(D') \in S] + \delta. \quad (51)$$

Therefore,

$$\Pr[(\phi\circ\mathcal{M})(D) \in E] \le e^\varepsilon \Pr[(\phi\circ\mathcal{M})(D') \in E] + \delta. \quad (52)$$

If ϕ is randomized independently of D, condition on its internal randomness and apply the deterministic argument to each realization. Averaging over the randomness of ϕ gives the same inequality. □

Lemma 7.3 (Sequential composition of RDP)**.** *Let $\mathcal{M}_1, \ldots, \mathcal{M}_K$ be a sequence of mechanisms, possibly chosen adaptively. Assume that for each i, conditioned on any fixed history of previous outputs, the conditional mechanism $\mathcal{M}_i$ satisfies (α, ρ_i)-RDP. Then the joint mechanism*

$$\mathcal{M}_{1:K}(D) = (\mathcal{M}_1(D), \ldots, \mathcal{M}_K(D)) \quad (53)$$

satisfies $(\alpha, \sum_{i=1}^K \rho_i)$-RDP.

Proof. Fix neighbouring datasets $D \sim D'$. Let P and Q be the joint laws of the transcript under D and D'. Write a transcript as $\mathbf{y}_{1:K} = (y_1, \ldots, y_K)$ and factor the likelihood ratio as

$$\frac{dP}{dQ}(\mathbf{y}_{1:K}) = \Pr\ od_{i=1}^K \frac{p_i(y_i \mid \mathbf{y}_{<i})}{q_i(y_i \mid \mathbf{y}_{<i})}, \quad (54)$$

Symbol	Meaning
D, D'	Neighbouring datasets differing in one record.
$\mathcal{M}$	A randomized mechanism.
ε, δ	Approximate-DP parameters.
α, ρ	RDP order and RDP parameter.
ρ_2	Order-2 RDP parameter.
$\mathbf{A}$	Frozen Gaussian factor in the compact PAVE sketch; its realization is internal in the intrinsic regime.
$\mathbf{B}(D)$	Released compact PAVE factor.
$\mathbf{F}(D)$	Dataset-dependent matrix query induced by the PAVE trajectory.
$\mathbf{U}$	Fixed public specification subspace with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$.
$\mathbf{H}(D)$	Coordinate representation of $\mathbf{F}(D)$ in the public subspace, so that $\mathbf{F}(D) = \mathbf{U}\mathbf{H}(D)$.
$\mathbf{C}(D)$	Coordinate compact release, $\mathbf{C}(D) = \mathbf{H}(D)\mathbf{A}^\top$.
$\mathbf{\Sigma}_D$	Gaussian-sketch covariance, $\mathbf{\Sigma}_D = \mathbf{H}(D)\mathbf{H}(D)^\top$.
r	Low-rank width of the Gaussian/PAVE factor.
k	Dimension of the public specification subspace.
γ	Relative covariance-stability parameter.
μ, L, Δ, G	Non-degeneracy, upper spectral scale, query sensitivity, and per-example bound used to verify covariance stability.
H	Baseline view in the risk analysis, e.g., model-only view. This is not the matrix $\mathbf{H}(D)$.
H_0	Baseline side information before observing the specification.
R	Released specification channel.
V	A generic released view in the binary neighbouring-dataset game.
$\mathcal{A}$	A possibly randomized adversary.
$p^V_{\mathcal{A}}(D)$	Probability that adversary $\mathcal{A}$ outputs 1 from view $V(D)$.
$\mathrm{Adv}^\star(V; D, D')$	Optimal binary distinguishing advantage between $V(D)$ and $V(D')$.
$\eta_{\varepsilon,\delta}$	Abbreviation for $(e^\varepsilon - 1) + \delta$.

Table 1: Notation used in the proofs. Bold uppercase symbols denote matrices. The baseline view H in the risk analysis is distinct from the matrix-valued query coordinate $\mathbf{H}(D)$.

where $p_i(\cdot \mid \mathbf{y}_{<i})$ and $q_i(\cdot \mid \mathbf{y}_{<i})$ are the conditional laws of the i-th output given the previous outputs under D and D', respectively. For any fixed history $\mathbf{y}_{<i}$, the conditional (α, ρ_i)-RDP assumption gives

$$\int \left(\frac{p_i(y_i \mid \mathbf{y}_{<i})}{q_i(y_i \mid \mathbf{y}_{<i})}\right)^{\alpha} q_i(dy_i \mid \mathbf{y}_{<i}) \le \exp((\alpha-1)\rho_i). \quad (55)$$

Therefore, by iterated expectation under Q,

$$\begin{aligned}\mathbb{E}_Q\left[\left(\frac{dP}{dQ}\right)^{\alpha}\right] &= \mathbb{E}_Q\left[\Pr\ od_{i=1}^{K}\left(\frac{p_i(Y_i \mid Y_{<i})}{q_i(Y_i \mid Y_{<i})}\right)^{\alpha}\right] \\ &\le \exp((\alpha-1)\rho_K)\mathbb{E}_Q\left[\Pr\ od_{i=1}^{K-1}\left(\frac{p_i(Y_i \mid Y_{<i})}{q_i(Y_i \mid Y_{<i})}\right)^{\alpha}\right] \\ &\le \cdots \le \exp\left((\alpha-1)\sum_{i=1}^{K}\rho_i\right).\end{aligned} \quad (56)$$

Taking logarithms and dividing by $\alpha - 1$ yields

$$D_\alpha(P\|Q) \le \sum_{i=1}^{K}\rho_i. \quad (57)$$

□

7.3 Gaussian Divergence Calculations

The compact PAVE release is a Gaussian sketch. The main calculation is the Rényi divergence between two centered Gaussians with different covariance matrices.

Lemma 7.4 (Rényi divergence between centered Gaussians)**.** *Let*

$$P = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \qquad Q = \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}'), \quad (58)$$

where $\mathbf{\Sigma}, \mathbf{\Sigma}' \in \mathbb{R}^{k\times k}$ are positive definite. Let

$$\mathbf{T} = (\mathbf{\Sigma}')^{-1/2}\mathbf{\Sigma}(\mathbf{\Sigma}')^{-1/2}, \quad (59)$$

and let $\lambda_1, \ldots, \lambda_k$ be the eigenvalues of $\mathbf{T}$. For $\alpha > 1$, if

$$\alpha + (1-\alpha)\lambda_i > 0 \qquad \text{for all } i, \quad (60)$$

then

$$D_\alpha(P\|Q) = \frac{1}{2(\alpha-1)}\sum_{i=1}^{k}\left[(1-\alpha)\log\lambda_i - \log(\alpha + (1-\alpha)\lambda_i)\right]. \quad (61)$$

In particular, for $\alpha = 2$, if $\lambda_i < 2$ for all i, then

$$D_2(P\|Q) = \frac{1}{2}\sum_{i=1}^{k} -\log(\lambda_i(2-\lambda_i)). \quad (62)$$

Proof. The densities of P and Q are

$$p(\mathbf{x}) = (2\pi)^{-k/2}|\mathbf{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{x}^\top\mathbf{\Sigma}^{-1}\mathbf{x}\right) \quad (63)$$

and

$$q(\mathbf{x}) = (2\pi)^{-k/2}|\mathbf{\Sigma}'|^{-1/2}\exp\left(-\frac{1}{2}\mathbf{x}^\top(\mathbf{\Sigma}')^{-1}\mathbf{x}\right). \quad (64)$$

For $\alpha > 1$,

$$\int p(\mathbf{x})^{\alpha} q(\mathbf{x})^{1-\alpha}\, d\mathbf{x} = |\mathbf{\Sigma}|^{-\alpha/2} |\mathbf{\Sigma}'|^{-(1-\alpha)/2} |\mathbf{K}|^{-1/2}, \tag{65}$$

where

$$\mathbf{K} = \alpha \mathbf{\Sigma}^{-1} + (1-\alpha)(\mathbf{\Sigma}')^{-1}. \tag{66}$$

The integral is finite exactly when $\mathbf{K} \succ 0$.

Now write

$$\mathbf{\Sigma} = (\mathbf{\Sigma}')^{1/2} \mathbf{T} (\mathbf{\Sigma}')^{1/2}. \tag{67}$$

Then

$$|\mathbf{\Sigma}| = |\mathbf{\Sigma}'|\, |\mathbf{T}| \tag{68}$$

and

$$\mathbf{K} = (\mathbf{\Sigma}')^{-1/2} \left[\alpha \mathbf{T}^{-1} + (1-\alpha)\mathbf{I}\right] (\mathbf{\Sigma}')^{-1/2}. \tag{69}$$

Thus

$$|\mathbf{K}| = |\mathbf{\Sigma}'|^{-1} \left|\alpha \mathbf{T}^{-1} + (1-\alpha)\mathbf{I}\right|. \tag{70}$$

Substituting these identities gives

$$\begin{aligned} &\log \int p^{\alpha} q^{1-\alpha} \\ &= -\frac{\alpha}{2} \log |\mathbf{\Sigma}| - \frac{1-\alpha}{2} \log |\mathbf{\Sigma}'| - \frac{1}{2} \log |\mathbf{K}| \\ &= -\frac{\alpha}{2} \log |\mathbf{T}| - \frac{1}{2} \log \left|\alpha \mathbf{T}^{-1} + (1-\alpha)\mathbf{I}\right|. \end{aligned} \tag{71}$$

Since the eigenvalues of $\mathbf{T}$ are $\lambda_1, \ldots, \lambda_k$,

$$\left|\alpha \mathbf{T}^{-1} + (1-\alpha)\mathbf{I}\right| = \Pr\ od_{i=1}^{k} \frac{\alpha + (1-\alpha)\lambda_i}{\lambda_i}. \tag{72}$$

Hence

$$\begin{aligned} &\log \int p^{\alpha} q^{1-\alpha} \\ &= -\frac{\alpha}{2} \sum_{i=1}^{k} \log \lambda_i - \frac{1}{2} \sum_{i=1}^{k} \log \frac{\alpha + (1-\alpha)\lambda_i}{\lambda_i} \\ &= \frac{1}{2} \sum_{i=1}^{k} \left[(1-\alpha) \log \lambda_i - \log(\alpha + (1-\alpha)\lambda_i)\right]. \end{aligned} \tag{73}$$

Dividing by $\alpha - 1$ gives Eq. (61).

For $\alpha = 2$,

$$\alpha + (1-\alpha)\lambda_i = 2 - \lambda_i. \tag{74}$$

Therefore

$$\begin{aligned} D_2(P\|Q) &= \frac{1}{2} \sum_{i=1}^{k} \left[-\log \lambda_i - \log(2-\lambda_i)\right] \\ &= \frac{1}{2} \sum_{i=1}^{k} -\log(\lambda_i(2-\lambda_i)). \end{aligned} \tag{75}$$

□

Lemma 7.5 (RDP of the Gaussian mechanism)**.** *Let $\mathbf{q}(D) \in \mathbb{R}^p$ be a vector-valued query with replacement-neighbour sensitivity*

$$\Delta_2 = \sup_{D \sim D'} \|\mathbf{q}(D) - \mathbf{q}(D')\|_2. \tag{76}$$

The Gaussian mechanism

$$\mathcal{G}(D) = \mathbf{q}(D) + \mathbf{Z}, \qquad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, s^2 \mathbf{I}_p), \tag{77}$$

satisfies (α, ρ_α)-RDP with

$$\rho_\alpha = \frac{\alpha \Delta_2^2}{2s^2}. \tag{78}$$

The same statement holds for matrix-valued queries after vectorization, with the Frobenius norm as the Euclidean norm.

Proof. For neighbouring $D \sim D'$, the two output laws are Gaussians with the same covariance:

$$P = \mathcal{N}(\mathbf{q}(D), s^2 \mathbf{I}_p), \qquad Q = \mathcal{N}(\mathbf{q}(D'), s^2 \mathbf{I}_p). \tag{79}$$

For Gaussians with the same covariance, the order-α Rényi divergence is

$$\begin{aligned} D_\alpha(P\|Q) &= \frac{\alpha}{2} (\mathbf{q}(D) - \mathbf{q}(D'))^{\top} (s^2 \mathbf{I}_p)^{-1} (\mathbf{q}(D) - \mathbf{q}(D')) \\ &= \frac{\alpha}{2s^2} \|\mathbf{q}(D) - \mathbf{q}(D')\|_2^2. \end{aligned} \tag{80}$$

The result follows by the sensitivity bound. For a matrix-valued query, vectorization preserves the Frobenius norm, so the same formula applies. □

7.4 Proof of the Compact Gaussian-Sketch DP Guarantee

We now prove Theorem 3.5. The key point is that the compact release

$$\mathbf{B}(D) = \mathbf{U}\mathbf{H}(D)\mathbf{A}^{\top} \tag{81}$$

is a matrix-normal Gaussian sketch once $\mathbf{A}$ is not released and the query $\mathbf{H}(D)$ is fixed conditional on the dataset.

Lemma 7.6 (Compact PAVE as a Gaussian sketch)**.** *Consider a single selected layer and omit the layer index. Let*

$$\mathbf{A} \in \mathbb{R}^{r \times m}, \qquad A_{ij} \sim \mathcal{N}(0, 1/r) \tag{82}$$

independently of D. Assume that, conditional on D, the PAVE query can be written as a deterministic matrix $\mathbf{F}(D) \in \mathbb{R}^{d \times m}$ that is independent of the realized $\mathbf{A}$. If

$$\mathbf{F}(D) = \mathbf{U}\mathbf{H}(D), \qquad \mathbf{U}^{\top}\mathbf{U} = \mathbf{I}_k, \tag{83}$$

then the compact release

$$\mathbf{B}(D) = \mathbf{F}(D)\mathbf{A}^{\top} \tag{84}$$

can be written as

$$\mathbf{B}(D) = \mathbf{U}\mathbf{C}(D), \qquad \mathbf{C}(D) = \mathbf{H}(D)\mathbf{A}^{\top}. \tag{85}$$

Moreover, the columns of $\mathbf{C}(D)$ are independent centered Gaussian vectors:

$$\begin{aligned} \mathbf{c}_j(D) &\sim \mathcal{N}\left(\mathbf{0}, \frac{1}{r}\mathbf{\Sigma}_D\right), \\ \mathbf{\Sigma}_D &= \mathbf{H}(D)\mathbf{H}(D)^{\top}, \qquad j = 1, \ldots, r. \end{aligned} \tag{86}$$

Proof. Write the rows of $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_r^\top \end{bmatrix}, \qquad \mathbf{a}_j \in \mathbb{R}^m. \tag{87}$$

Since $A_{ij} \sim \mathcal{N}(0, 1/r)$ independently, each row transpose satisfies

$$\mathbf{a}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m/r), \tag{88}$$

and $\mathbf{a}_1, \dots, \mathbf{a}_r$ are independent.

If $\mathbf{F}(D) = \mathbf{U}\mathbf{H}(D)$, then

$$\mathbf{B}(D) = \mathbf{F}(D)\mathbf{A}^\top = \mathbf{U}\mathbf{H}(D)\mathbf{A}^\top = \mathbf{U}\mathbf{C}(D). \tag{89}$$

The j-th column of $\mathbf{C}(D)$ is

$$\mathbf{c}_j(D) = \mathbf{H}(D)\mathbf{a}_j. \tag{90}$$

Since $\mathbf{H}(D)$ is deterministic conditional on D and independent of $\mathbf{a}_j$, a linear transformation of a centered Gaussian vector gives

$$\begin{aligned} \mathbf{c}_j(D) &\sim \mathcal{N}\left(\mathbf{0}, \mathbf{H}(D)\frac{\mathbf{I}_m}{r}\mathbf{H}(D)^\top\right) \\ &= \mathcal{N}\left(\mathbf{0}, \frac{1}{r}\mathbf{H}(D)\mathbf{H}(D)^\top\right). \end{aligned} \tag{91}$$

The independence of the columns follows from the independence of $\mathbf{a}_1, \dots, \mathbf{a}_r$. □

Theorem 7.7 (Intrinsic DP for compact PAVE specifications)**.** *Consider the compact PAVE mechanism*

$$\mathcal{M}_{\mathbf{B}}(D) = \mathbf{B}(D) = \mathbf{U}\mathbf{H}(D)\mathbf{A}^\top, \tag{92}$$

where $A_{ij} \sim \mathcal{N}(0, 1/r)$ is sampled independently of D, and the realized $\mathbf{A}$ is not released. Assume that $\mathbf{U} \in \mathbb{R}^{d\times k}$ is fixed and public with $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_k$. Let

$$\mathbf{\Sigma}_D = \mathbf{H}(D)\mathbf{H}(D)^\top. \tag{93}$$

Assume that for every ordered neighbouring pair $D \sim D'$,

$$(1-\gamma)\mathbf{\Sigma}_{D'} \Pr\ eceq \mathbf{\Sigma}_D \Pr\ eceq (1+\gamma)\mathbf{\Sigma}_{D'}, \qquad 0 < \gamma < 1, \tag{94}$$

and that the covariance matrices are positive definite. Then $\mathcal{M}_{\mathbf{B}}$ satisfies $(2, \rho_2)$-RDP with

$$\rho_2 \le \frac{rk}{2}\log\frac{1}{1-\gamma^2}. \tag{95}$$

Consequently, for every $\delta \in (0, 1)$, $\mathcal{M}_{\mathbf{B}}$ is (ε, δ)-DP with

$$\varepsilon = \frac{rk}{2}\log\frac{1}{1-\gamma^2} + \log\frac{1}{\delta}. \tag{96}$$

Moreover,

$$\varepsilon \le \frac{rk\gamma^2}{2(1-\gamma^2)} + \log\frac{1}{\delta}. \tag{97}$$

Proof. By Lemma 7.6,

$$\mathbf{B}(D) = \mathbf{U}\mathbf{C}(D), \qquad \mathbf{C}(D) = \mathbf{H}(D)\mathbf{A}^\top. \tag{98}$$

Since $\mathbf{U}$ is fixed and public, $\mathbf{B}(D)$ is a deterministic post-processing of $\mathbf{C}(D)$. It suffices to bound the Rényi divergence between $\mathbf{C}(D)$ and $\mathbf{C}(D')$.

Write

$$\mathbf{C}(D) = [\mathbf{c}_1(D), \dots, \mathbf{c}_r(D)]. \tag{99}$$

For each j,

$$\mathbf{c}_j(D) \sim \mathcal{N}\left(\mathbf{0}, \frac{1}{r}\mathbf{\Sigma}_D\right), \tag{100}$$

and the columns are independent. For a single column, define

$$\begin{aligned} P_D &= \mathcal{N}\left(\mathbf{0}, \frac{1}{r}\mathbf{\Sigma}_D\right), \\ P_{D'} &= \mathcal{N}\left(\mathbf{0}, \frac{1}{r}\mathbf{\Sigma}_{D'}\right). \end{aligned} \tag{101}$$

The common scale factor $1/r$ cancels in the covariance ratio. Let

$$\mathbf{T} = \mathbf{\Sigma}_{D'}^{-1/2}\mathbf{\Sigma}_D\mathbf{\Sigma}_{D'}^{-1/2}. \tag{102}$$

By covariance stability,

$$(1-\gamma)\mathbf{I}_k \Pr\ eceq \mathbf{T} \Pr\ eceq (1+\gamma)\mathbf{I}_k. \tag{103}$$

Hence each eigenvalue λ_i of $\mathbf{T}$ satisfies

$$1 - \gamma \le \lambda_i \le 1 + \gamma. \tag{104}$$

Since $\gamma < 1$, we have $\lambda_i < 2$ for all i. By Lemma 7.4 with $\alpha = 2$,

$$D_2(P_D \| P_{D'}) = \frac{1}{2}\sum_{i=1}^{k} -\log(\lambda_i(2-\lambda_i)). \tag{105}$$

For every i,

$$\lambda_i(2-\lambda_i) = 1 - (\lambda_i - 1)^2. \tag{106}$$

Since $|\lambda_i - 1| \le \gamma$,

$$\lambda_i(2-\lambda_i) \ge 1 - \gamma^2. \tag{107}$$

Thus

$$D_2(P_D \| P_{D'}) \le \frac{k}{2}\log\frac{1}{1-\gamma^2}. \tag{108}$$

The full coordinate release $\mathbf{C}(D)$ contains r independent columns. Therefore its law is $P_D^{\otimes r}$, while the law of $\mathbf{C}(D')$ is $P_{D'}^{\otimes r}$. Rényi divergence is additive over independent product distributions, so

$$\begin{aligned} D_2(\mathbf{C}(D) \| \mathbf{C}(D')) &= r\, D_2(P_D \| P_{D'}) \\ &\le \frac{rk}{2}\log\frac{1}{1-\gamma^2}. \end{aligned} \tag{109}$$

By post-processing, the same $(2, \rho_2)$-RDP bound holds for $\mathbf{B}(D) = \mathbf{U}\mathbf{C}(D)$. Applying Lemma 7.1 with $\alpha = 2$ gives Eq. (96).

Finally, for $0 \le x < 1$,

$$\log\frac{1}{1-x} \le \frac{x}{1-x}. \tag{110}$$

Taking $x = \gamma^2$ yields Eq. (97). □

7.5 Verifying Covariance Stability

The main theorem uses covariance stability. The next two lemmas justify the sufficient conditions discussed in the main text.

Lemma 7.8 (Sensitivity implies covariance stability)**.** *Suppose that for all admissible datasets D,*

$$\mu^2 \mathbf{I}_k \Pr\ eceq \mathbf{H}(D)\mathbf{H}(D)^\top \Pr\ eceq L^2 \mathbf{I}_k, \tag{111}$$

and for all neighbouring datasets $D \sim D'$,

$$\|\mathbf{H}(D) - \mathbf{H}(D')\|_2 \le \Delta. \tag{112}$$

Let $\mathbf{\Sigma}_D = \mathbf{H}(D)\mathbf{H}(D)^\top$. Then

$$\|\mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'}\|_2 \le 2L\Delta. \tag{113}$$

Moreover, if

$$\gamma = \frac{2L\Delta}{\mu^2} < 1, \tag{114}$$

then

$$(1-\gamma)\mathbf{\Sigma}_{D'} \Pr\ eceq \mathbf{\Sigma}_D \Pr\ eceq (1+\gamma)\mathbf{\Sigma}_{D'}. \tag{115}$$

Proof. Let

$$\mathbf{H} = \mathbf{H}(D), \qquad \mathbf{H}' = \mathbf{H}(D'). \tag{116}$$

Then

$$\mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'} = \mathbf{H}\mathbf{H}^\top - \mathbf{H}'(\mathbf{H}')^\top. \tag{117}$$

We decompose

$$\mathbf{H}\mathbf{H}^\top - \mathbf{H}'(\mathbf{H}')^\top = (\mathbf{H}-\mathbf{H}')\mathbf{H}^\top + \mathbf{H}'(\mathbf{H}-\mathbf{H}')^\top. \tag{118}$$

Thus

$$\begin{aligned} \|\mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'}\|_2 &\le \|(\mathbf{H}-\mathbf{H}')\mathbf{H}^\top\|_2 \\ &\quad + \|\mathbf{H}'(\mathbf{H}-\mathbf{H}')^\top\|_2 \\ &\le (\|\mathbf{H}\|_2 + \|\mathbf{H}'\|_2)\|\mathbf{H}-\mathbf{H}'\|_2. \end{aligned} \tag{119}$$

The upper spectral bound in Eq. (111) implies $\|\mathbf{H}(D)\|_2 \le L$ for all admissible D. Therefore Eq. (113) follows.

Equivalently,

$$-2L\Delta\, \mathbf{I}_k \Pr\ eceq \mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'} \Pr\ eceq 2L\Delta\, \mathbf{I}_k. \tag{120}$$

Since $\mathbf{\Sigma}_{D'} \succeq \mu^2 \mathbf{I}_k$,

$$\mathbf{I}_k \Pr\ eceq \frac{1}{\mu^2}\mathbf{\Sigma}_{D'}. \tag{121}$$

Hence

$$-\frac{2L\Delta}{\mu^2}\mathbf{\Sigma}_{D'} \Pr\ eceq \mathbf{\Sigma}_D - \mathbf{\Sigma}_{D'} \Pr\ eceq \frac{2L\Delta}{\mu^2}\mathbf{\Sigma}_{D'}. \tag{122}$$

This is exactly Eq. (115). □

Lemma 7.9 (Sensitivity of an average-type query)**.** *Suppose*

$$\mathbf{H}(D) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{\Psi}(z_i), \tag{123}$$

where

$$\|\mathbf{\Psi}(z)\|_2 \le G \tag{124}$$

for all records z. If $D \sim D'$ differ in exactly one record, then

$$\|\mathbf{H}(D) - \mathbf{H}(D')\|_2 \le \frac{2G}{n}. \tag{125}$$

Proof. Let

$$D = \{z_1, \dots, z_n\}, \qquad D' = \{z_1', \dots, z_n'\}, \tag{126}$$

and suppose the two datasets differ only at index j. Then

$$\mathbf{H}(D) - \mathbf{H}(D') = \frac{1}{n}\left(\mathbf{\Psi}(z_j) - \mathbf{\Psi}(z_j')\right). \tag{127}$$

Therefore

$$\begin{aligned} \|\mathbf{H}(D) - \mathbf{H}(D')\|_2 &\le \frac{1}{n}\left(\|\mathbf{\Psi}(z_j)\|_2 + \|\mathbf{\Psi}(z_j')\|_2\right) \\ &\le \frac{2G}{n}. \end{aligned} \tag{128}$$

□

7.6 Why the Realized Gaussian Factor Must Remain Internal

We next prove Proposition 3.7. This result justifies the compact secret-release condition in the intrinsic theorem. The sampling distribution of $\mathbf{A}$ may be public, but the realized matrix cannot be released in the intrinsic compact-sketch regime.

Proposition 7.10 (A public-$\mathbf{A}$ compact release is not generally DP)**.** *Consider the mechanism*

$$\mathcal{M}_{\mathrm{pub}}(D) = (\mathbf{A}, \mathbf{F}(D)\mathbf{A}^\top), \tag{129}$$

where $\mathbf{A} \in \mathbb{R}^{r\times m}$ has i.i.d. Gaussian entries and the realized $\mathbf{A}$ is released. Suppose there exist neighbouring datasets $D \sim D'$ such that

$$\mathbf{F}(D) - \mathbf{F}(D') \neq \mathbf{0}. \tag{130}$$

Then, for this neighbouring pair, $\mathcal{M}_{\mathrm{pub}}$ is not (ε, δ)-DP for any finite ε and any $\delta < 1$.

Proof. Let

$$\mathbf{L} = \mathbf{F}(D) - \mathbf{F}(D'). \tag{131}$$

By assumption, $\mathbf{L} \neq \mathbf{0}$. Define the event

$$E_D = \left\{(\mathbf{a}, \mathbf{b}) : \mathbf{b} = \mathbf{F}(D)\mathbf{a}^\top\right\}. \tag{132}$$

Under dataset D, the mechanism outputs

$$\mathcal{M}_{\mathrm{pub}}(D) = (\mathbf{A}, \mathbf{F}(D)\mathbf{A}^\top), \tag{133}$$

so

$$\Pr\left[\mathcal{M}_{\mathrm{pub}}(D) \in E_D\right] = 1. \tag{134}$$

Under dataset D', the output lies in E_D if and only if

$$\mathbf{F}(D')\mathbf{A}^\top = \mathbf{F}(D)\mathbf{A}^\top, \tag{135}$$

or equivalently

$$\mathbf{L}\mathbf{A}^\top = \mathbf{0}. \tag{136}$$

Write the rows of $\mathbf{A}$ as $\mathbf{a}_1^\top, \dots, \mathbf{a}_r^\top$. Then Eq. (136) means

$$\mathbf{L}\mathbf{a}_j = \mathbf{0} \qquad \text{for all } j = 1, \dots, r. \tag{137}$$

Because $\mathbf{L} \neq \mathbf{0}$, its nullspace is a proper linear subspace of $\mathbb{R}^m$. A non-degenerate Gaussian vector lies in a proper linear subspace with probability zero. Therefore,

$$\Pr\left[\mathbf{L}\mathbf{a}_j = \mathbf{0}\right] = 0 \tag{138}$$

for each j, and hence

$$\Pr\left[\mathbf{L}\mathbf{A}^\top = \mathbf{0}\right] = 0. \tag{139}$$

Thus

$$\Pr\left[\mathcal{M}_{\mathrm{pub}}(D') \in E_D\right] = 0. \tag{140}$$

If $\mathcal{M}_{\mathrm{pub}}$ were (ε, δ)-DP, applying DP to the event E_D would give

$$1 \le e^{\varepsilon} \cdot 0 + \delta = \delta, \tag{141}$$

which is impossible for $\delta < 1$. Hence no finite ε and no $\delta < 1$ can satisfy DP for this neighbouring pair. □

Remark 7.11 (Expanded Wishart releases are a different mechanism)**.** *The intrinsic theorem analyzes the compact release $\mathbf{F}(D)\mathbf{A}^\top$ with the realized $\mathbf{A}$ kept internal. It does not analyze the expanded update $\mathbf{F}(D)\mathbf{A}^\top\mathbf{A}$. The expanded update is a quadratic, Wishart-type release and can have algebraic support constraints that are absent from the compact Gaussian sketch.*

For example, in dimension two, take

$$\mathbf{F} = \mathbf{I}_2, \qquad \mathbf{F}' = \begin{pmatrix} 1+\alpha & 0 \\ 0 & 1 \end{pmatrix}, \qquad \alpha > 0. \tag{142}$$

Then $\mathbf{F}\mathbf{A}^\top\mathbf{A}$ is symmetric almost surely. In contrast, $\mathbf{F}'\mathbf{A}^\top\mathbf{A}$ is symmetric only if $\mathbf{A}^\top\mathbf{A}$ commutes with $\mathbf{F}'$, which requires the off-diagonal entry of $\mathbf{A}^\top\mathbf{A}$ to be zero. For Gaussian $\mathbf{A}$, this event has probability zero. Thus the compact-release theorem should not be read as a theorem about expanded Wishart releases.

7.7 Multi-Layer and Shared-Randomness Releases

The main theorem is stated for a single compact Gaussian sketch. A learnware specification may combine multiple layers, and several specifications may share dock-internal randomness. The following results justify the release-pattern discussion in the main text.

Lemma 7.12 (Multi-layer composition under independent layer randomness)**.** *For each selected layer $\ell \in \mathcal{L}$, suppose the compact release*

$$\mathbf{B}_\ell(D) = \mathbf{U}_\ell\mathbf{H}_\ell(D)\mathbf{A}_\ell^\top \tag{143}$$

satisfies $(2, \rho_{2,\ell})$-RDP, and suppose the layer-wise Gaussian matrices $\{\mathbf{A}_\ell\}_{\ell\in\mathcal{L}}$ are sampled independently. Then the joint release

$$\mathcal{M}_{\mathcal{L}}(D) = \{\mathbf{B}_\ell(D)\}_{\ell\in\mathcal{L}} \tag{144}$$

satisfies $(2, \rho_{2,\mathcal{L}})$-RDP with

$$\rho_{2,\mathcal{L}} = \sum_{\ell\in\mathcal{L}} \rho_{2,\ell}. \tag{145}$$

In particular, if layer ℓ satisfies Theorem 3.5 with parameters $(r_\ell, k_\ell, \gamma_\ell)$, then

$$\rho_{2,\mathcal{L}} \le \sum_{\ell\in\mathcal{L}} \frac{r_\ell k_\ell}{2} \log\frac{1}{1-\gamma_\ell^2}. \tag{146}$$

Proof. This is a special case of RDP composition. Because the layer-wise Gaussian matrices are independent conditional on D, the joint law is a product law:

$$P = \bigotimes_{\ell\in\mathcal{L}} P_\ell, \qquad Q = \bigotimes_{\ell\in\mathcal{L}} Q_\ell, \tag{147}$$

where P_ℓ and Q_ℓ are the laws of $\mathbf{B}_\ell(D)$ and $\mathbf{B}_\ell(D')$. For order 2,

$$D_2(P\|Q) = \log\int \left(\frac{dP}{dQ}\right)^2 dQ. \tag{148}$$

Since the likelihood ratio factorizes,

$$\frac{dP}{dQ} = \Pr\ od_{\ell\in\mathcal{L}}\frac{dP_\ell}{dQ_\ell}, \tag{149}$$

we obtain

$$\begin{aligned}\int \left(\frac{dP}{dQ}\right)^2 dQ &= \int \Pr\ od_{\ell\in\mathcal{L}} \left(\frac{dP_\ell}{dQ_\ell}\right)^2 \Pr\ od_{\ell\in\mathcal{L}} dQ_\ell \\ &= \Pr\ od_{\ell\in\mathcal{L}} \int \left(\frac{dP_\ell}{dQ_\ell}\right)^2 dQ_\ell.\end{aligned} \tag{150}$$

Taking logarithms yields

$$D_2(P\|Q) = \sum_{\ell\in\mathcal{L}} D_2(P_\ell\|Q_\ell) \le \sum_{\ell\in\mathcal{L}} \rho_{2,\ell}. \tag{151}$$

The layer-wise bound follows from Theorem 7.7. □

Lemma 7.13 (Stacked release under shared dock randomness)**.** *Suppose J compact specifications are generated using the same internal Gaussian matrix $\mathbf{A}$:*

$$\mathbf{B}_j(D) = \mathbf{U}_j\mathbf{H}_j(D)\mathbf{A}^\top, \qquad j = 1, \dots, J. \tag{152}$$

Define the stacked release and the stacked public subspace by

$$\bar{\mathbf{B}}(D) = \begin{bmatrix} \mathbf{B}_1(D) \\ \vdots \\ \mathbf{B}_J(D) \end{bmatrix}, \qquad \bar{\mathbf{U}} = \mathrm{diag}(\mathbf{U}_1, \dots, \mathbf{U}_J). \tag{153}$$

Define the stacked coordinate query by

$$\bar{\mathbf{H}}(D) = \begin{bmatrix} \mathbf{H}_1(D) \\ \vdots \\ \mathbf{H}_J(D) \end{bmatrix}. \tag{154}$$

Then

$$\bar{\mathbf{B}}(D) = \bar{\mathbf{U}}\bar{\mathbf{H}}(D)\mathbf{A}^\top. \tag{155}$$

If

$$\bar{\mathbf{\Sigma}}_D = \bar{\mathbf{H}}(D)\bar{\mathbf{H}}(D)^\top \tag{156}$$

is positive definite and satisfies

$$(1-\bar\gamma)\bar{\mathbf{\Sigma}}_{D'} \Pr\ eceq \bar{\mathbf{\Sigma}}_D \Pr\ eceq (1+\bar\gamma)\bar{\mathbf{\Sigma}}_{D'}, \qquad 0 < \bar\gamma < 1, \tag{157}$$

then the stacked release satisfies $(2, \bar\rho_2)$-RDP with

$$\bar\rho_2 \le \frac{r\bar{k}}{2}\log\frac{1}{1-\bar\gamma^2}, \tag{158}$$

where $\bar{k}$ is the row dimension of $\bar{\mathbf{H}}(D)$.

Proof. The stacked release has exactly the same form as the single compact release:

$$\bar{\mathbf{B}}(D) = \bar{\mathbf{U}}\bar{\mathbf{H}}(D)\mathbf{A}^\top. \tag{159}$$

The matrix $\bar{\mathbf{U}}$ is fixed and public. The coordinate release

$$\bar{\mathbf{C}}(D) = \bar{\mathbf{H}}(D)\mathbf{A}^\top \tag{160}$$

has r independent Gaussian columns, each with covariance

$$\frac{1}{r}\bar{\mathbf{\Sigma}}_D = \frac{1}{r}\bar{\mathbf{H}}(D)\bar{\mathbf{H}}(D)^\top. \tag{161}$$

The proof of Theorem 7.7 applies verbatim with

$$(\mathbf{H}, \mathbf{\Sigma}, k, \gamma) \quad \text{replaced by} \quad (\bar{\mathbf{H}}, \bar{\mathbf{\Sigma}}, \bar{k}, \bar{\gamma}). \tag{162}$$

This gives the stated RDP bound. □

7.8 Proof of DP-S-PAVE

We now prove the certified DP guarantee for DP-S-PAVE. The proof is standard DP-SGD/RDP accounting applied in the low-rank PAVE update space.

Lemma 7.14 (Sensitivity of a clipped projected update)**.** *Let*

$$\mathrm{Clip}(\mathbf{Z}, C) = \mathbf{Z} \cdot \min\left\{1, \frac{C}{\|\mathbf{Z}\|_{\mathrm{F}}}\right\}. \tag{163}$$

For any two per-example projected updates $\mathbf{Z}$ *and* $\mathbf{Z}'$*,*

$$\|\mathrm{Clip}(\mathbf{Z}, C)\|_{\mathrm{F}} \le C, \tag{164}$$

and

$$\|\mathrm{Clip}(\mathbf{Z}, C) - \mathrm{Clip}(\mathbf{Z}', C)\|_{\mathrm{F}} \le 2C. \tag{165}$$

Consequently, for replacement-neighbour datasets, the sum of clipped projected updates has Frobenius sensitivity at most $2C$*.*

Proof. The first inequality follows directly from the definition of clipping. For the second,

$$\begin{aligned} &\|\mathrm{Clip}(\mathbf{Z}, C) - \mathrm{Clip}(\mathbf{Z}', C)\|_{\mathrm{F}} \\ &\quad \le \|\mathrm{Clip}(\mathbf{Z}, C)\|_{\mathrm{F}} + \|\mathrm{Clip}(\mathbf{Z}', C)\|_{\mathrm{F}} \\ &\quad \le 2C. \end{aligned} \tag{166}$$

If two datasets differ by replacing one record, all clipped terms in the mini-batch sum cancel except the contribution of the changed record. Thus the sum changes by at most $2C$ in Frobenius norm. □

Theorem 7.15 (Certified DP of DP-S-PAVE)**.** *Fix the matrices* $\{\mathbf{A}_\ell\}_{\ell\in\mathcal{L}}$ *independently of the dataset. In DP-S-PAVE, assume that every per-example projected update is clipped to Frobenius norm at most* C*, and independent Gaussian noise with entrywise standard deviation* σC *is added before deterministic scaling. Let* $\rho_\alpha^{\mathrm{subG}}(q, \sigma)$ *denote the order-*α *RDP parameter of one Poisson-subsampled Gaussian update with replacement-neighbour sensitivity* $2C$ *and noise standard deviation* σC*. Then the final release*

$$R = \{\mathbf{B}_{\ell,T}\}_{\ell\in\mathcal{L}} \tag{167}$$

satisfies (α, ρ_α)*-RDP with*

$$\rho_\alpha \le \sum_{t=0}^{T-1} \sum_{\ell\in\mathcal{L}} \rho_\alpha^{\mathrm{subG}}(q, \sigma). \tag{168}$$

Consequently, for every $\delta \in (0, 1)$*, DP-S-PAVE is* $(\varepsilon_{\mathrm{dp}}, \delta)$*-DP with*

$$\varepsilon_{\mathrm{dp}}(\delta) = \inf_{\alpha>1}\left\{\rho_\alpha + \frac{\log(1/\delta)}{\alpha - 1}\right\}. \tag{169}$$

If $q = 1$*, then a conservative closed-form bound is*

$$\rho_\alpha \le \frac{2\alpha T|\mathcal{L}|}{\sigma^2}, \tag{170}$$

and

$$\varepsilon_{\mathrm{dp}} \le \frac{2T|\mathcal{L}|}{\sigma^2} + 2\sqrt{\frac{2T|\mathcal{L}|}{\sigma^2}\log\frac{1}{\delta}}. \tag{171}$$

Proof. Fix the matrices $\{\mathbf{A}_\ell\}_{\ell\in\mathcal{L}}$ and condition on the entire past history before step t. This includes previous noisy updates, current parameters, and all public randomness. Conditioned on this history, the per-example projected update at layer ℓ is a deterministic function of the current record, the current model state, and the fixed $\mathbf{A}_\ell$. After clipping, Lemma 7.14 shows that the replacement-neighbour sensitivity of the sum of clipped projected updates is at most $2C$ in Frobenius norm.

The update at layer ℓ and step t adds a Gaussian matrix $\mathbf{N}_{\ell,t}$ with independent entries from $\mathcal{N}(0, \sigma^2C^2)$. After vectorization, this is a Gaussian mechanism with Euclidean sensitivity $2C$ and noise standard deviation σC, possibly preceded by Poisson subsampling with rate q. By definition of $\rho_\alpha^{\mathrm{subG}}(q, \sigma)$, this conditional update satisfies

$$(\alpha, \rho_\alpha^{\mathrm{subG}}(q, \sigma))\text{-RDP}. \tag{172}$$

The deterministic scaling by the batch-size constant, the gradient step with learning rate η, and the subsequent update of $\mathbf{B}_{\ell,t}$ are post-processing operations and do not change the RDP parameter.

There are $T|\mathcal{L}|$ such conditional Gaussian updates. By adaptive RDP composition, Lemma 7.3, the final release satisfies Eq. (168). The conversion to $(\varepsilon_{\mathrm{dp}}, \delta)$-DP follows from Lemma 7.1 and optimizing over $\alpha > 1$.

It remains to derive the closed-form full-batch bound. When $q = 1$, there is no subsampling. By Lemma 7.5, one Gaussian update has RDP parameter

$$\rho_{\alpha,\mathrm{one}} = \frac{\alpha(2C)^2}{2(\sigma C)^2} = \frac{2\alpha}{\sigma^2}. \tag{173}$$

Composing over $T|\mathcal{L}|$ updates gives Eq. (170). Let

$$a = \frac{2T|\mathcal{L}|}{\sigma^2}, \qquad L_\delta = \log\frac{1}{\delta}. \tag{174}$$

Then the RDP-to-DP conversion gives

$$\varepsilon(\alpha) = a\alpha + \frac{L_\delta}{\alpha - 1}. \tag{175}$$

Writing $\beta = \alpha - 1 > 0$,

$$\varepsilon(\alpha) = a + a\beta + \frac{L_\delta}{\beta}. \tag{176}$$

The minimum over $\beta > 0$ is attained at

$$\beta = \sqrt{\frac{L_\delta}{a}}, \tag{177}$$

which yields Eq. (171). □

Remark 7.16 (Post-processing consequences of DP-S–PAVE)**.** *Once DP-S-PAVE is applied, the DP guarantee comes from explicit Gaussian perturbation in the low-rank update space. Therefore the compact factors $\mathbf{B}_\ell$, the expanded updates $\mathbf{B}_\ell\mathbf{A}_\ell$, normalized specifications, similarity scores, rankings, and top-K retrieval outputs are all post-processing of a DP mechanism. By Lemma 7.2, they inherit the same DP guarantee. This is why DP-S-PAVE can cover deployments where $\mathbf{A}$ or $\mathbf{BA}$ must be public, whereas the intrinsic compact-sketch theorem cannot.*

7.9 Binary Gain and Distinguishing Advantage

The main text defines general gain-based privacy risks, and then analyzes their binary neighbouring-dataset instantiation through distinguishing advantage. We record the precise relationship.

For a released view V and a binary adversary $\mathcal{A}$, define

$$p^V_{\mathcal{A}}(D) = \Pr\left[\mathcal{A}(V(D)) = 1\right]. \tag{178}$$

The optimal distinguishing advantage is

$$\mathrm{Adv}^\star(V; D, D') = \sup_{\mathcal{A}} \left| p^V_{\mathcal{A}}(D) - p^V_{\mathcal{A}}(D') \right|. \tag{179}$$

Lemma 7.17 (Binary gain and advantage)**.** *Consider the balanced binary game in which the challenger samples $b \sim \mathrm{Unif}\{0,1\}$, releases $V(D)$ if $b = 1$ and $V(D')$ if $b = 0$, and the adversary outputs $\hat{b}$. The optimal success probability satisfies*

$$\mathrm{gain}^\star_{\mathrm{bin}}(V; D, D') = \frac{1}{2} + \frac{1}{2}\,\mathrm{Adv}^\star(V; D, D'). \tag{180}$$

Proof. For a fixed adversary $\mathcal{A}$ that outputs 1 when it guesses $b = 1$,

$$\begin{aligned}\mathrm{gain}_{\mathrm{bin}}(\mathcal{A}; V, D, D') &= \frac{1}{2}\Pr\left[\mathcal{A}(V(D)) = 1\right] \\ &\quad + \frac{1}{2}\Pr\left[\mathcal{A}(V(D')) = 0\right] \\ &= \frac{1}{2} + \frac{1}{2}\left(p^V_{\mathcal{A}}(D) - p^V_{\mathcal{A}}(D')\right).\end{aligned} \tag{181}$$

If the difference is negative, the adversary can swap its output labels and obtain the absolute value. Taking the supremum over all adversaries gives Eq. (180). □

7.10 Proofs of Learnware Risk Guarantees

We now prove Lemma 4.1, Theorem 4.2, and Theorem 4.3. The central idea is that, after fixing a baseline view, any attack using the additional released specification channel is a post-processing of that channel.

Lemma 7.18 (DP channels add bounded inference advantage)**.** *Let H be a baseline view, and let $R_D = \mathcal{M}(D)$ be an additional released channel. Assume that R is conditionally (ε, δ)-DP given H in the following sense: for a chosen version of the regular conditional law of R_D given $H(D) = u$, the inequality*

$$\begin{aligned}&\Pr\left[R_D \in S \mid H(D) = u\right] \le \\ &e^\varepsilon \Pr\left[R_{D'} \in S \mid H(D') = u\right] + \delta\end{aligned} \tag{182}$$

holds for every fixed baseline value u, every measurable event S, and all neighbouring datasets $D \sim D'$ for which the conditional distributions are considered. Then

$$\mathrm{Adv}^\star((H, R); D, D') \le \mathrm{Adv}^\star(H; D, D') + (e^\varepsilon - 1) + \delta \tag{183}$$

for all neighbouring datasets $D \sim D'$.

Remark 7.19 (On the conditional-DP assumption)**.** *The conditional-DP assumption in Lemma 7.18 is essential. If the baseline view H is itself data-dependent, ordinary DP of R does not automatically imply conditional DP of R given H. The assumption is satisfied, for example, when H is fixed public side information, or when the specification mechanism satisfies the same DP guarantee uniformly after conditioning on each admissible value of the model-only view. This is the formal condition under which the specification-side amplification theorem separates leakage already present in H from the extra contribution of R.*

Proof. Fix a binary adversary $\mathcal{A}$ that acts on the joint view (H, R). For a fixed baseline value u, define

$$g_D(u) = \Pr\left[\mathcal{A}(u, R_D) = 1 \mid H(D) = u\right]. \tag{184}$$

By conditional DP and post-processing,

$$g_D(u) \le e^\varepsilon g_{D'}(u) + \delta. \tag{185}$$

Since $g_{D'}(u) \le 1$,

$$g_D(u) \le g_{D'}(u) + (e^\varepsilon - 1) + \delta. \tag{186}$$

Let

$$\eta_{\varepsilon,\delta} = (e^\varepsilon - 1) + \delta. \tag{187}$$

Averaging over the distribution of $H(D)$ gives

$$\begin{aligned}p^{(H,R)}_{\mathcal{A}}(D) &= \mathbb{E}_{H(D)}[g_D(H)] \\ &\le \mathbb{E}_{H(D)}[g_{D'}(H)] + \eta_{\varepsilon,\delta}.\end{aligned} \tag{188}$$

The map $u \mapsto g_{D'}(u)$ is a randomized binary test that uses only the baseline view H. Therefore, by the definition of $\mathrm{Adv}^\star(H; D, D')$,

$$\mathbb{E}_{H(D)}[g_{D'}(H)] \le \mathbb{E}_{H(D')}[g_{D'}(H)] + \mathrm{Adv}^\star(H; D, D'). \tag{189}$$

But

$$\mathbb{E}_{H(D')}[g_{D'}(H)] = p^{(H,R)}_{\mathcal{A}}(D'). \tag{190}$$

Combining Eqs. (188)–(190) yields

$$p^{(H,R)}_{\mathcal{A}}(D) - p^{(H,R)}_{\mathcal{A}}(D') \le \mathrm{Adv}^\star(H; D, D') + \eta_{\varepsilon,\delta}. \tag{191}$$

Repeating the same argument with D and D' exchanged gives

$$p^{(H,R)}_{\mathcal{A}}(D') - p^{(H,R)}_{\mathcal{A}}(D) \le \mathrm{Adv}^\star(H; D, D') + \eta_{\varepsilon,\delta}. \tag{192}$$

Thus

$$\left|p_{\mathcal{A}}^{(H,R)}(D) - p_{\mathcal{A}}^{(H,R)}(D')\right| \le \mathrm{Adv}^{\star}(H; D, D') + \eta_{\varepsilon,\delta}. \tag{193}$$

Taking the supremum over all binary adversaries $\mathcal{A}$ proves the lemma. □

Theorem 7.20 (DP controls specification disclosure risk)**.** *Let $R = \mathcal{M}_{\mathbf{B}}(D)$ be the released compact PAVE specification. If R is conditionally (ε, δ)-DP given the baseline side information H_0, then the advantage-style disclosure risk satisfies*

$$\mathrm{Risk}_{\mathrm{dis}}^{\mathrm{adv}}(R \mid H_0) \le (e^{\varepsilon} - 1) + \delta. \tag{194}$$

If H_0 is fixed public side information and R satisfies ordinary (ε, δ)-DP, the same bound holds.

Proof. By the definition of advantage-style disclosure risk,

$$\mathrm{Risk}_{\mathrm{dis}}^{\mathrm{adv}}(R \mid H_0) = \sup_{D \sim D'} \Big[\mathrm{Adv}^{\star}((H_0, R); D, D') - \mathrm{Adv}^{\star}(H_0; D, D') \Big]_+. \tag{195}$$

Applying Lemma 7.18 with $H = H_0$ gives, for every neighbouring pair,

$$\mathrm{Adv}^{\star}((H_0, R); D, D') - \mathrm{Adv}^{\star}(H_0; D, D') \le (e^{\varepsilon} - 1) + \delta. \tag{196}$$

Taking the positive part and the supremum over neighbouring pairs gives the desired bound.

If H_0 is fixed public side information, conditioning on H_0 does not change the distribution of R in a data-dependent way. Thus ordinary DP of R implies conditional DP given H_0, and the same argument applies. □

Theorem 7.21 (No material specification-side amplification)**.** *Let $R = \mathcal{M}_{\mathbf{B}}(D)$ be the released compact PAVE specification, and let H be the model-only view. If R is conditionally (ε, δ)-DP given H, then the advantage-style amplification risk satisfies*

$$\mathrm{Risk}_{\mathrm{amp}}^{\mathrm{adv}}(H; R) \le (e^{\varepsilon} - 1) + \delta. \tag{197}$$

Proof. By definition,

$$\mathrm{Risk}_{\mathrm{amp}}^{\mathrm{adv}}(H; R) = \sup_{D \sim D'} \Big[\mathrm{Adv}^{\star}((H, R); D, D') - \mathrm{Adv}^{\star}(H; D, D') \Big]_+. \tag{198}$$

Applying Lemma 7.18 with the model-only view H as the baseline gives, for every neighbouring pair,

$$\mathrm{Adv}^{\star}((H, R); D, D') - \mathrm{Adv}^{\star}(H; D, D') \le (e^{\varepsilon} - 1) + \delta. \tag{199}$$

Taking the positive part and the supremum over neighbouring pairs proves the theorem. □

Remark 7.22 (Interpreting the two risk bounds)**.** *The disclosure and amplification bounds have the same numerical margin because in both cases the newly added channel is the same DP-protected specification channel R. Their meanings are different. For disclosure, the baseline is side information H_0, and the question is what the specification reveals by itself. For amplification, the baseline is the model-only view H, and the question is whether the specification gives extra distinguishing power beyond what the model already reveals. The proofs differ only in the choice of baseline view in Lemma 7.18.*

7.11 Additional Consequences

We end by recording two consequences that are used implicitly in the main text.

Corollary 7.23 (Binary gain increase)**.** *If*

$$\mathrm{Risk}_{\mathrm{dis}}^{\mathrm{adv}}(R \mid H_0) \le \eta, \tag{200}$$

then in the corresponding balanced binary disclosure game, the optimal gain increase from H_0 to (H_0, R) is at most $\eta/2$. Similarly, if

$$\mathrm{Risk}_{\mathrm{amp}}^{\mathrm{adv}}(H; R) \le \eta, \tag{201}$$

then in the corresponding balanced binary amplification game, the optimal gain increase from H to (H, R) is at most $\eta/2$.

Proof. By Lemma 7.17, the optimal binary gain equals

$$\frac{1}{2} + \frac{1}{2}\,\mathrm{Adv}^{\star}(\cdot). \tag{202}$$

Thus replacing an advantage by another advantage larger by at most η increases the optimal binary gain by at most $\eta/2$. □

Corollary 7.24 (Risk guarantees under DP-S-PAVE)**.** *If the released specification channel is produced by DP-S-PAVE with privacy parameters $(\varepsilon_{\mathrm{dp}}, \delta)$, then*

$$\mathrm{Risk}_{\mathrm{dis}}^{\mathrm{adv}}(R \mid H_0) \le (e^{\varepsilon_{\mathrm{dp}}} - 1) + \delta \tag{203}$$

whenever R is conditionally DP given H_0, and

$$\mathrm{Risk}_{\mathrm{amp}}^{\mathrm{adv}}(H; R) \le (e^{\varepsilon_{\mathrm{dp}}} - 1) + \delta \tag{204}$$

whenever R is conditionally DP given the model-only view H.

Proof. Theorem 7.15 gives the DP parameters of the released specification channel. Substituting these parameters into Theorem 7.20 and Theorem 7.21 gives the two bounds. □

7.12 Summary of the Proof Chain

The proof chain established above is as follows. First, the compact PAVE release

$$\mathbf{B}(D) = \mathbf{U}\mathbf{H}(D)\mathbf{A}^{\top} \tag{205}$$

is a Gaussian sketch of the projected matrix query $\mathbf{H}(D)$ when the realized $\mathbf{A}$ is kept internal and the query is independent of that realization. Second, covariance stability ensures

that the neighbouring Gaussian sketch distributions have finite and controlled Rényi divergence:

$$D_2(\mathcal{M}_{\mathbf{B}}(D)\|\mathcal{M}_{\mathbf{B}}(D')) \leq \frac{rk}{2}\log\frac{1}{1-\gamma^2}. \tag{206}$$

Third, RDP-to-DP conversion gives the compact intrinsic DP guarantee. Fourth, if the intrinsic conditions are not enforced, DP-S-PAVE provides a certified DP guarantee by clipping per-example low-rank updates and adding calibrated Gaussian noise. Finally, once the released specification channel is DP, Lemma 7.18 converts this mechanism-level guarantee into bounds on disclosure risk and specification-side amplification risk.