THE WISHART PROJECTION MECHANISM IS ALMOST DIFFERENTIALLY PRIVATE: GUARANTEES AND LIMITATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce and analyse the *Wishart projection mechanism*, a randomised map of the form $\mathcal{A}_r(S) := f(S)M$ where $M \sim \mathsf{W}_d(\frac{1}{r}I_d, r)$, and study its differential privacy (DP) properties. For vector-valued queries f, we prove non-asymptotic DP guarantees showing that Wishart randomness alone suffices, without requiring any additive noise. For matrix-valued queries, however, we establish a sharp separation: we prove that without additive noise the mechanism is not DP, and empirically supplement this theoretical result with a near-perfect membership inference attack ($\mathrm{AUC} > 0.99$). As a concrete instantiation, we show that LoRA-style updates are an instance of the matrix-valued Wishart projection mechanism, implying that LoRA is not inherently private despite its initialisation randomness, and that empirical observations of reduced memorisation under LoRA should not be interpreted as DP guarantees. We then analyse a noisy variant and prove that the Wishart projection can *amplify* the resulting privacy guarantee: at a fixed noise multiplier, composing with a freshly resampled projection yields strictly stronger DP bounds than noise alone, both in the large-r and small-r regimes.

1 INTRODUCTION

Differential Privacy (DP) (Dwork et al., 2006) is widely regarded as the gold standard for protecting the privacy of individuals represented in training data. At a high level, DP limits the influence of any individual training example on the model's output, thereby making it difficult to determine whether a particular example was included in the training set. Differential privacy is typically achieved by introducing randomness through additive noise. Yet randomness is already ubiquitous in many machine learning algorithms, arising from mechanisms such as random projections, sketches, and random initialization. This raises a natural question: can the randomness already present in an algorithm provide meaningful privacy guarantees, without the need to introduce additional noise?

Blocki et al. (2012) showed that the Johnson-Lindenstrauss transform itself preserves privacy: a Gaussian sketch of a well-conditioned matrix is (ε, δ)-DP with no additive noise at all. In this paper we ask if a similar result holds when we use another class of projections, namely Wishart matrices. This family arises naturally from low-rank adaptation (Hu et al., 2022): under LoRA's standard initialisation $B_0 = 0$, the gradient with respect to A vanishes at the first step, so the first weight update is exactly $W_1 - W_0 = -\eta \nabla_W \mathcal{L}(W_0)\, A_0^\top A_0$, the gradient right-multiplied by a rank-r Wishart matrix. We formalize this through the following abstraction

Definition 1 (Projection mechanism)**.** *Let $\mathcal{S} \subset \mathcal{X}^n$ be a dataset collection and $f : \mathcal{S} \to \mathbb{R}^{n\times d}$ a query function. For $r \in \mathbb{N}$, the (Wishart) projection mechanism is defined by*

$$\mathcal{A}_r(S) \coloneqq f(S)M, \qquad M := Z^\top Z,$$

where $Z \in \mathbb{R}^{r\times d}$ has i.i.d rows $z_k \overset{i.i.d}{\sim} \mathcal{N}\left(0, \frac{1}{r}I_d\right)$ or equivalently, $M \sim \mathsf{W}_d(\frac{1}{r}I_d,\, r)$[1].

For vector-valued queries $f : \mathcal{S} \to \mathbb{R}^d$, we establish non-asymptotic differential privacy guarantees for the vector-valued projection mechanism in Theorem 1. One might expect the matrix-valued

[1]All of our analysis extends to Gaussian noise $Z \sim \mathcal{N}(0, \sigma^2 I_d)$ for any $\sigma^2 > 0$. We set $\sigma^2 = 1/r$ to match the usual normalization of Wishart random projection (e.g. in LoRA initializations).

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

case to follow with a dimension-dependent degradation. It does not. For matrix-valued queries $f : \mathcal{S} \to \mathbb{R}^{n\times d}$ the same mechanism, satisfies no (ε, δ) guarantee for *any* ε and any $\delta < 1$ (Proposition 2). The failure is not quantitative but geometric. In the full-rank case every output satisfies the identity $V^\top Y = Y^\top V$, which holds almost surely for the true input but with probability zero for a non-aligned neighbour. Thus, the symmetry of M imposes a dataset-dependent algebraic constraint on the output, and the constraint sets corresponding to neighbouring inputs have measure-zero overlap. This is also what separates us from Blocki et al. (2012), whose Gaussian randomness has full support on the linear image of the sketching map. The consequence for LoRA is immediate: without additional additive noise, neighbouring datasets can yield non-overlapping support, ruling out any (finite-ε, finite-δ) DP guarantee. This negative result is empirically supported under a weaker threat model, where a membership inference attack achieves near-perfect performance ($\mathrm{AUC} \geq 0.99$).

Support collapse indicates that some form of additive noise is necessary to obtain meaningful privacy guarantees. We therefore consider a noisy variant of the projection mechanism, obtained by adding a small amount of additive noise. This enforces overlap between the output distributions of neighbouring datasets, and the above attack no longer succeeds. In this setting we extend our guarantees from vectors to matrices and show that Wishart projections can amplify privacy: for fixed noise, composing with the projection yields strictly stronger privacy bounds than the noise level alone would suggest, both for large r (Section 4.1) and in the practically relevant small-r regime (Section 4.2). A key implication is that low rank is beneficial beyond utility: at a fixed noise multiplier, projection can amplify privacy and yield smaller ε than full fine-tuning, a point not made explicit in prior DP-LoRA analyses.

Our Contributions (i) We introduce the Wishart projection mechanism and prove non-asymptotic DP guarantees in the vector case. (ii) We mathematically show the corresponding matrix mechanism is not DP, and demonstrate its vulnerability via a near-perfect MIA ($\mathrm{AUC} > 0.99$). (iii) For a variant involving additive noise, we prove privacy amplification for both large and small ranks.

Organization. In Section 2 we review DP preliminaries and give a brief overview of why LoRA is an instance of the Wishart projection mechanism. Section 3 proves privacy of the Wishart projection mechanism for vector-valued functions and shows why it fails for matrices. Section 4 analyses the noisy Wishart projection mechanism for matrix-valued functions and establishes privacy amplification in both rank regimes. Finally, Section 5 covers related work, limitations, and open questions.

2 Preliminaries

Differential privacy limits how much the output distribution can change when a single data point is modified. We say datasets S, S' are neighbors, denoted $S \sim_H S'$, if they differ in a single entry (i.e., Hamming distance $d_H(S, S') = 1$).

Definition 2 (Differential privacy)**.** *A randomised algorithm $\mathcal{A} : \mathcal{S} \to \mathcal{Y}$ is (ε, δ)-DP if for all measurable $E \subseteq \mathcal{Y}$ and all $S \sim_H S'$,*

$$\Pr(\mathcal{A}(S) \in E) \leq e^{\varepsilon} \Pr(\mathcal{A}(S') \in E) + \delta,$$

with probability taken over the internal randomness of $\mathcal{A}$.

Differential privacy is commonly enforced by additive perturbations: adding noise to the output of a non-private query f, calibrated to its ℓ_2 sensitivity, $\Delta := \max_{S\sim_H S'} \|f(S) - f(S')\|_2$.

Lemma 1 (Gaussian Mechanism)**.** *Let $f : \mathcal{S} \to \mathbb{R}^d$ be a function with ℓ_2-sensitivity Δ. For any $\varepsilon, \delta \in (0, 1)$, the Gaussian mechanism $\mathcal{A}(S) = f(S) + Z$ is (ε, δ)-DP, for*

$$Z \sim \mathcal{N}(0, \sigma^2 I_d), \text{ where } \sigma \geq \frac{2\Delta\sqrt{\log(1.25/\delta)}}{\varepsilon}.$$

Restricting $\mathcal{A}$ to datasets in $\mathcal{D} \subseteq \mathcal{S}$ gives DP guarantees *conditioned on* $\mathcal{D}$. This often reduces Δ and improves utility, but offers no privacy outside $\mathcal{D}$. One way to obtain full privacy is to privately check whether the dataset belongs to $\mathcal{D}$, e.g., via Propose-Test-Release Dwork & Lei (2009).

LoRA is a Projection Mechanism Low-Rank Adaptation (LoRA; Hu et al. (2022)) is one of the most popular parameter-efficient fine-tuning approaches for Large Language Models. It augments a

pretrained weight matrix $W_0 \in \mathbb{R}^{n\times d}$ by a *low-rank* update:

$$W = W_0 + BA,\ B \in \mathbb{R}^{n\times r},\ A \in \mathbb{R}^{r\times d},\ r \ll \min\{d, n\}.$$

By freezing W_0 and only optimizing the smaller matrices (B, A), LoRA reduces computational cost while preserving the base model's capabilities.

LoRA-FA(Zhang et al., 2023) is a variant that fixes A and only updates B. When A is initialized Gaussian and then *frozen*, we update B by $B_{t+1} = B_t - \eta\,(\nabla_W \mathcal{L}(W_t))A^\top$. Then after T steps, we can write

$$W_T = W_0 - \eta \sum_{t=1}^{T} \left(\nabla_W \mathcal{L}(W_t)\right)(A^\top A). \tag{1}$$

Thus each step uses a gradient $\nabla_W \mathcal{L}(W_t) \in \mathbb{R}^{n\times d}$ that is *right-projected* by the random matrix $A^\top A$ (a rank-r Wishart). Right multiplying a vector or matrix by a randomly sampled Wishart is precisely the projection mechanism studied in this work. Prior work (Malekmohammadi & Farnadi, 2024) suggested that this random projection structure may yield DP guarantees. However, as we show in Section 3, the matrix-valued projection mechanism is not DP in general (see Proposition 2).

However, our results in Section 4 show that noisy versions of the projection mechanism enjoy (ε, δ)-DP guarantees that in certain regimes are stronger than simple noise addition (i.e., Gaussian mechanism) would suggest due to the randomness stemming from the projection.

3 Projection Mechanism: Privacy and Limits

We study the privacy properties of the projection mechanism (Definition 1). We first analyze the special case of vector-valued queries $f : \mathcal{S} \to \mathbb{R}^d$ and show it is differentially private under an alignment assumption (Section 3.1). In contrast, for matrix-valued queries we prove a negative result: Proposition 2 shows $\varepsilon = \infty$ for any $\delta < 1$, which implies that LoRA-FA and LoRA is not private.

3.1 Privacy Guarantees for Vector-Valued Functions

Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a query with $\|f(S)\|_2 = 1$ for all S; this normalization can always be enforced by scaling, and we assume it throughout. For a dataset collection $\mathcal{D} \subset \mathcal{X}^n$, define the *minimum alignment*

$$\rho(\mathcal{D}, f) \coloneqq \min_{\substack{S,S'\in\mathcal{D}\\ S\sim_H S'}} f(S)^\top f(S') \in [-1, 1]. \tag{2}$$

When the context is clear, we write $\rho := \rho(\mathcal{D}, f)$. Large ρ means that neighbouring query outputs are nearly co-directional. In this regime the laws of $Mf(S)$ and $Mf(S')$ are harder to distinguish, leading to tighter privacy guarantees (Figure 1). Crucially, much like a sensitivity parameter in additive mechanisms, ρ is a property fixed by the dataset collection $\mathcal{D}$ and the query f and is not controlled by the projection mechanism itself. One can also preprocess the data to improve the alignment ρ through e.g. noise addition. See Appendix C.1 for details.

Let $t_\ell(\cdot)$ and $\kappa_\ell(\cdot)$ denote the quantile functions of the Student-t and χ^2 distributions, respectively, each with ℓ degrees of freedom.

Theorem 1. *For a dataset collection $\mathcal{D}$ and a query function f with outputs in $\mathbb{R}^d$ and $\|f(S)\|_2 = 1$ for all $S \in \mathcal{D}$, let $\rho > 0$ be the minimum alignment defined in Equation* (2)*. Let $\eta \in (0, 1/3)$, $\nu = d + r - 1$, and*

$$K = \sqrt{\frac{1-\rho^2}{r}}\, t_r(1-\eta), \quad \Delta = \max\{\nu - \kappa_\nu(\eta/2), \kappa_\nu(1-\eta/2) - \nu\}.$$

If $\rho > K$, then the projection mechanism (Definition 1) with rank r is $(\varepsilon, 3\eta)$-DP on $\mathcal{D}$ for

$$\varepsilon \geq \frac{\nu + \Delta}{2}\left(\frac{1}{\rho - K} + \log(\rho - K) - 1\right) - (r - 1 + \Delta/2)\log(\rho - K).$$

Here r is the rank of the Wishart matrix $M = ZZ^\top$ with $Z \in \mathbb{R}^{d\times r}$, and is an algorithmic choice. As illustrated in Figure 1 (left), for fixed δ the resulting ε depends non-monotonically on r. This reflects

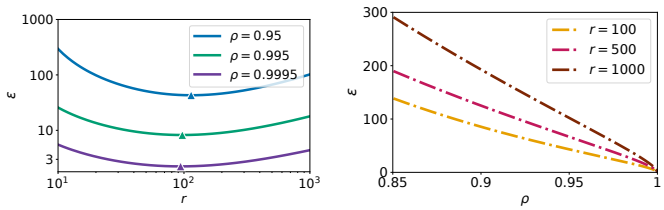


Figure 1: Privacy bound from Theorem 1 at $\delta = 10^{-3}$ and $d = 1024$.

two competing effects: smaller r makes M lower-rank, confining the output to a lower-dimensional subspace and potentially reducing leakage via compression; but it also averages over fewer Gaussian directions, reducing the concentration of the privacy loss random variable.

On the dependence of ρ on n. The minimum alignment ρ measures how much the direction of the normalised query can change under a single record change. For average-type queries, one replacement perturbs the unnormalised vector by $O(1/n)$, which typically yields $\rho = 1 - \widetilde{O}(1/n^2)$. Concretely, let $\tilde{f}(S) = \frac{1}{n}\sum_{i=1}^{n} g(x_i)$ with $\|g(x)\|_2 \le L$ and assume $\inf_{S\in\mathcal{D}} \|\tilde{f}(S)\|_2 \ge c_0 > 0$. Then for neighboring datasets S, S', we have $\|\tilde{f}(S) - \tilde{f}(S')\|_2 \le 2L/n$. Defining $f(S) := \tilde{f}(S)/\|\tilde{f}(S)\|_2$ gives $\|f(S) - f(S')\|_2 \le 4L/(c_0 n)$ and hence $\rho \ge 1 - \frac{8L^2}{c_0^2 n^2}$, using $1 - u^\top v = \frac{1}{2}\|u - v\|_2^2$ for unit vectors. The same argument applies to mini-batch gradients: for $\tilde{f}(B) = \frac{1}{|B|}\sum_{x\in B} \nabla\ell(\theta; x)$, if $\|\nabla\ell(\theta; x)\|_2 \le L$ and $\|\tilde{f}(B)\|_2$ is bounded away from zero, then $\rho = 1 - \widetilde{O}(1/|B|^2)$.

Some direct applications of the Wishart Mechanism that applies Theorem 1, including projected gradient descent and private retrieval, are discussed in Section C.2.

3.2 Negative Results for Matrix-Valued Functions

Theorem 1 establishes (ε, δ)-DP for the projection mechanism on *vector*-valued queries under an alignment condition. Since the LoRA-FA update can be cast as the same mechanism with *matrix*-valued outputs (see Section 2), one might expect an analogous guarantee for matrices and thus privacy for LoRA-FA. We show this is false: without additive noise, the matrix-valued projection mechanism is not DP..

Proposition 2. *Let $f : \mathcal{S} \to \mathbb{R}^{n\times d}$ be any non-trivial function and $\mathcal{A}(S)$ be the projection mechanism (Definition 1) with query function f. Then, there exists two neighboring datasets $S, S' \in \mathcal{S}$ and a measurable event E such that*

$$\Pr(\mathcal{A}(S) \in E) = 1 \quad \textit{and} \quad \Pr(\mathcal{A}(S') \in E) = 0.$$

Consequently, $\mathcal{A}$ is not (ε, δ)-DP for any ε and any $\delta < 1$.

Proof intuition: By non-trivial we mean that f can produce neighboring outputs $V = f(S)$ and $V' = f(S')$ for which either V has a kernel direction a with $V'a \neq 0$, or V has full column rank and V' is not a scalar multiple of V. For matrix-valued queries $f : \mathcal{S} \to \mathbb{R}^{n\times d}$ with $n, d > 1$, this is a mild condition: any query that can change in a genuinely matrix-valued way between neighboring datasets satisfies it. In the rank-deficient case, $Va = 0$, so every output $Y = ZZ^\top V$ satisfies $Ya = 0$, whereas $Z'Z'^\top V'a \neq 0$ almost surely when $V'a \neq 0$. In the full-rank case, every output satisfies that $V^\top Y = V^\top ZZ^\top V$ is symmetric, while for non-aligned V' this identity holds only on the zero set of a nonzero polynomial, hence with probability zero. Thus some measurable event E has probability 1 under one neighboring dataset and 0 under the other, violating (ε, δ)-DP for every $\delta < 1$ (see Lemma 13).

Setting	Metric	Rank r		
		16	**128**	**512**
Scratch	Acc. (%)	57.31	66.36	68.36
	AUC (%)	99.86	100.00	100.00
Pre.+FT	Acc. (%)	75.98	76.12	76.45
	AUC (%)	99.81	100.00	100.00

Table 1: MIA performance for noise-free LoRA under a static-poison canary construction.

Implication for LoRA. Under LoRA's standard initialisation $B_0 = 0$ the gradient with respect to A vanishes at the first step, so $W_1 - W_0 = -\eta\nabla_W \mathcal{L}(W_0)A_0^\top A_0$ is an instance of Definition 1 with $M = A_0^\top A_0$ drawn independently of the

data. Releasing W_1 is therefore not (ε, δ)-DP for any ε and any $\delta < 1$ (See Appendix D.1 for detailed explanation). This extends to any release containing W_1, in particular the transcript $(W_1, \ldots, W_T)$ that DP-SGD-style accounting assumes. If only the final weights W_T are published our impossibility result does not directly apply, we therefore complement the theory with evidence in exactly that regime: only having access to the final fine-tuned model, a membership inference adversary with a static-poison canary achieves $\mathrm{AUC} > 0.99$ (Table 1).

We run 2000 independent trials on CIFAR-10 using the popular static-poison canary construction Nasr et al. (2021) in two regimes: (i) training from scratch, and (ii) pretraining followed by fine-tuning on CIFAR-10. Table 1 reports the model's test accuracy and ROC-AUC of the MIA. Across both regimes and all ranks, the attack achieves near-perfect separability (AUC ≈ 1), consistent with the non-overlap intuition suggested by the theory. See Appendix I.1 for details.

These findings do not conflict with prior work showing that LoRA can reduce MIA success (Hong et al., 2025), since MIA performance depends strongly on the threat model. We consider a strong adversary who knows all but one training records and can choose an adversarial canary, mirroring DP's worst-case auxiliary-information setting where the guarantee must hold for any choice of differing example. Our work shows that under a classical DP style adversary, LoRA is not private.

4 The Noisy Projection Mechanism

As shown in the previous section, the symmetry constraint induced by a Wishart (PSD) projection can make the output supports of neighboring matrix inputs disjoint. This motivates adding a small amount of Gaussian noise to enforce overlap if we want to keep DP guarantee.

In this section, we show that once overlap exists, the intrinsic randomness of the Wishart projection can amplify privacy beyond what an analysis based on additive noise alone would suggest. We study two noisy variants (M1 and M2), and distinguish two complementary amplification mechanisms: (i) a large-rank phenomenon (Section 4.1) driven by posterior-stable residual randomness in M; and (ii) a small-rank phenomenon (Section 4.2) arising from sensitivity contraction under projection onto a low-dimensional random subspace.

Setup and Mechanism. Let $d, n, r \geq 1$ and fix the noise scale $\sigma_G > 0$. Draw $Z \in \mathbb{R}^{d\times r}$ with i.i.d. entries $Z_{ik} \sim \mathcal{N}(0, 1/r)$ and set $M := ZZ^\top \in \mathbb{R}^{d\times d}$. Let $\Xi := [\xi_1, \ldots, \xi_n] \in \mathbb{R}^{d\times n}$ with i.i.d columns $\xi_k \sim \mathcal{N}(0, \sigma_G^2 I_d)$.

Definition 3 (Noisy Projection Mechanisms)**.** *Given $V = [v_1, \ldots, v_n] \in \mathbb{R}^{d\times n}$, define[2]:*

$$\mathcal{A}_1(V) := MV + \Xi \tag{M1}$$

$$\mathcal{A}_2(V) := M(V + \Xi) \tag{M2}$$

4.1 Matrix Projection in the large r regime

We first study the mechanism $\mathcal{A}_1(V) = MV + \Xi$ (M1). Unlike the vector case in Section 3, privacy can fail even when only one column of V changes. The issue is that the same random matrix M is reused across columns: after observing $n-1$ released columns, an adversary can learn information about M and use it to infer the remaining column more accurately. Thus, some additive noise Ξ is needed to mask this leakage and ensure distributional overlap.

To prove the privacy of M1, we first prove a privacy guarantee for a one-column adjacency, and then lift it to general matrix adjacency by composing the one-columns steps. In this section, we assume all inputs columns are unit vectors. For $V, V' \in \mathbb{R}^{d\times n}$, write $J(V, V') := \{j \in [n] : v_j \neq v'_j\}$, and for any $j \in [n]$, let V_{-j} denote V with column j removed. Now, we define the adjacency relations.

Definition 4 (Adjacency notion)**.** *Fix parameters $\beta_\perp \in [0, 1]$ and $\rho_\parallel, \rho_\perp \in [-1, 1]$.*

1. ***(One-column adjacency).*** *Fix $j \in [n]$ and set $S := \mathrm{span}(V_{-j}) \subset \mathbb{R}^d$. We write $V \sim^{(j)} V'$ if $V_{-j} = V'_{-j}$ and, writing $v_j = a + b$ and $v'_j = a' + b'$ with $a, a' \in S$ and $b, b' \in S^\perp$, we have*

$$\|b\|_2, \|b'\|_2 = \beta_\perp,\ \langle a, a'\rangle \geq \rho_\parallel \left(1 - \beta_\perp^2\right),\ \langle b, b'\rangle \geq \rho_\perp \beta_\perp^2$$

[2]We adopt left-multiplication in M1 and M2 for analytical convenience, but they have the identical privacy guarantee as Definition 1.

2. ***(Matrix adjacency).*** *Write $V \sim V'$ if there exists an ordering $j_1, \dots, j_k$ of changed columns $J(V, V')$, matrices $V = V^{(0)}, \dots, V^{(k)} = V'$ such that $V^{(t-1)} \sim^{(j_t)} V^{(t)}$ for all $t \in [k]$.*

The parameter $\beta_\perp$ measures how much of the changed column lies outside the span of the unchanged input columns. The parameters $\rho_\parallel$ and $\rho_\perp$ control alignment within S and $S^\perp$, respectively.

Theorem 3 (Privacy of M1 in the large-r regime)**.** *Fix $\beta_\perp, \rho_\parallel, \rho_\perp$ and let $V \sim V'$ with $k = |J(V, V')|$. Define $(\varepsilon_{\text{one}}, \delta_{\text{one}})$ as the privacy parameters for a single admissible step $V \sim^{(j)} V'$ (given in Theorem 4). Then $\mathcal{A}_1$ is (ε, δ)-DP for inputs V, V' for all $j \in [k]$, with*

$$\varepsilon \le k\,\varepsilon_{\text{one}}, \qquad \delta \le k\,e^{(k-1)\varepsilon_{\text{one}}}\,\delta_{\text{one}}.$$

4.1.1 Proof Sketch

Let $V^{(0)}, \dots, V^{(k)}$ be the adjacency chain from Definition 4. As in group privacy, it suffices to prove a privacy guarantee for a single step where only one column changes, and then use a group-privacy style analysis to compose those guarantees along the chain. So the main work is the one-column case.

Fix a step $V \sim^{(j)} V'$, so $V_{-j} = V'_{-j}$. Split the mechanism's output as:

$$(X, Y) \coloneqq (Mv_j + \xi_j,\ MV_{-j} + \Xi_{-j}),$$

where X is the release of the changed column and Y contains the release of the unchanged columns.

Let $S = \text{span}(V_{-j})$, let U have orthonormal columns spanning S, and set $G = U^\top Z$. The decomposition $M = M_\parallel + M_\perp$ from Appendix E.1 satisfies $M_\perp V_{-j} = 0$, so $Y = M_\parallel V_{-j} + \Xi_{-j}$. Moreover, conditional on G, the residual $M_\perp$ is independent of Y. Thus, observing the unchanged-column releases does not alter the conditional distribution of the residual. This allows us to apply the vector-valued privacy bound to the residual contribution of the changed columns.

Decompose $v_j = a + b$ and $v'_j = a' + b'$ with $a, a' \in S$ and $b, b' \in S^\perp$. Since $M_\perp a = 0$, the released column splits as

$$Mv_j + \xi_j = \underbrace{(M_\parallel v_j + \xi_j)}_{=:C(v_j)} + \underbrace{M_\perp b}_{=:R(b)}.$$

Residual term. For $R(b)$, the *vector-valued* DP analysis in Section 3, provides a $(\varepsilon_\perp, \delta_\perp)$ bound.

Lemma 2 (Residual DP bound)**.** *Let $s \coloneqq \dim(S)$ and $q \coloneqq r - p$, and $\nu = d - s + q - 1$ where $p = \text{rank}(G)$. Fix $\rho_\perp \in (0, 1]$, $\delta' \in (0, 1/3)$, and set $K_\perp = t_q(1-\delta')\sqrt{1-\rho_\perp^2/q}, \Delta_\perp = \max\{\nu - k_\nu(\delta'/2), k_\nu(1-\delta'/2) - \nu\}$. Then, if $\rho_\perp > K_\perp$, $R(b) = M_\perp b$ is $(\varepsilon_\perp, 3\delta')$-DP for the residual inputs in Definition 4 where*

$$\varepsilon_\perp = \frac{\nu + \Delta_\perp}{2}\left(\frac{1}{\rho_\perp - K_\perp} + \log(\rho_\perp - K_\perp) - 1\right) - (q - 1 + \Delta_\perp/2)\log(\rho_\perp - K_\perp).$$

Correlated term. For the correlated term $C(v_j)$, we apply the Gaussian mechanism using the directional sensitivity bound in Lemma 15. For a neighbouring pair, set $u = (v_j - v'_j)/\|v_j - v'_j\|$. With probability at least $1 - \delta_s$, $\|M_\parallel(v_j - v'_j)\|_2 \le \Gamma_{\delta_s}(u)\|v_j - v'_j\|_2$. The adjacency conditions give

$$\|v_j - v'_j\|_2^2 \le \Delta_v^2 \coloneqq 2\Big[(1-\beta_\perp^2)(1-\rho_\parallel) + \beta_\perp^2(1-\rho_\perp)\Big]. \tag{3}$$

Conditioning on this event we get the correlated term is $(\varepsilon_\parallel, \delta_\parallel) - DP$ with

$$\varepsilon_\parallel \coloneqq \frac{\Gamma_{\delta_s}\Delta_v}{\sigma_G}\sqrt{2\ln(1.25/\delta_\parallel)}. \tag{4}$$

Finally, sequential composition gives $\varepsilon_{\text{one}} \le \varepsilon_\parallel + \varepsilon_\perp$ and $\delta_{\text{one}} \le \delta_\parallel + \delta_\perp + \delta_s$.

Theorem 4. *Fix $j \in [n]$. Let $V \sim^{(j)} V'$ with parameters $(\rho_\parallel, \beta_\perp, \rho_\perp)$ defined in Definition 4. Fix $\delta_\parallel, \delta_s \in (0, 1)$, and choose $\delta' > 0$ as in Lemma 2 to obtain $(\varepsilon_\perp, \delta_\perp)$. Let $p = \text{rank}(U^\top Z)$ and define Γ_{δ_s} as above, and $\varepsilon_\parallel$ as in Equation (4). Then the mechanism $\mathcal{A}_1(V) = MV + \Xi$ is $(\varepsilon_{\text{one}}, \delta_{\text{one}})$-DP for inputs $V \sim_A^{(j)} V'$, with $\varepsilon_{\text{one}} \le \varepsilon_\parallel + \varepsilon_\perp$ and $\delta_{\text{one}} \le \delta_\parallel + \delta_\perp + \delta_s$.*

Composing the one-column bounds along the chain yields Theorem 3.

Comparison to the Gaussian mechanism. Standard Gaussian accounting is conservative in that it calibrates to a worst-case sensitivity $\|M\|\|\Delta V\|_F$ for all M, whereas our accounting takes into account the randomness of M and can be tighter when the dataset collection satisfies favorable geometric conditions on $\beta_\perp$, $\rho_\perp$, and $\rho_\parallel$. In particular, the Gaussian baseline gives $\varepsilon^{(k)}_{\text{class}} \lesssim \sigma_M^2(\sqrt{d}+\sqrt{r})^2\sqrt{k}\,K_G$, with $K_G := \frac{\Delta_v}{\sigma_G}\sqrt{2\ln(1.25/\delta_G)}$ and $\sigma_M^2 = 1/r$ a comparison with our bound yields

$$\frac{\varepsilon^{(k)}_{\text{ours}}}{\varepsilon^{(k)}_{\text{class}}} \lesssim \underbrace{\frac{\sqrt{rk(1-\beta_\perp^2)(1-\rho_\parallel)+pk\beta_\perp^2(1-\rho_\perp)}}{(\sqrt{d}+\sqrt{r})\Delta_v}}_{\text{Correlated part}} + \underbrace{\frac{\sqrt{k}\varepsilon_\perp}{\sigma_M^2(\sqrt{d}+\sqrt{r})^2 K_G}}_{\text{Residual part}},$$

where $p = \operatorname{rank}(U^\top Z) \le s$ is the effective dimension of the revealed column span. When $\rho_\perp$ is close to 1 and σ_G is small, the residual part is negligible. We then focus on $\beta_\perp > \sqrt{1-1/r}$, where the changed columns lie mostly outside the span of the unchanged columns. If the $\rho_\perp$ and $\rho_\parallel$ are of comparable order, the residual-direction term dominates and the correlated numerator has dimensional dependence $\sqrt{pk}$. Since $r \gg d \ge p$, this can be substantially smaller than the $\sqrt{r}+\sqrt{d}$ scaling of the Gaussian baseline, yielding a tighter privacy bound. Figure 2 illustrates this regime for two choices of σ_G; see Section F for derivation of the comparison and Section H for details on the dataset construction.

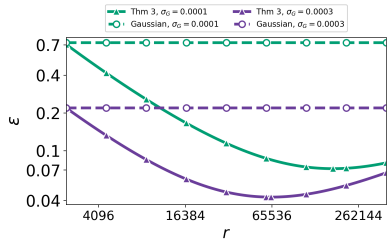


Figure 2: Large r v.s. Gaussian

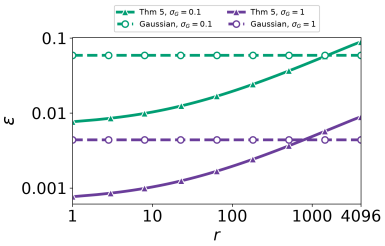


Figure 3: Small r v.s. Gaussian

4.2 Matrix Projection in the Small r Regime

The small-r regime yields privacy amplification through a different mechanism from the large r regime. Here we analyse $\mathcal{A}_2(V) = M(V+\Xi)$ (M2). One might naively treat M as post-processing: add Gaussian noise Ξ calibrated to the Frobenius sensitivity $\|\Delta V\|_F$, then multiply by M. However, the important observation is that the output of M2 always lies in the random column space $\operatorname{col}(M) = \operatorname{col}(Z)$. Consequently, conditioned on M, privacy depends only on the projected difference $P_M \Delta V$, where P_M is the orthogonal projector onto $\operatorname{col}(M)$. When $r \ll d$ the projection typically captures only a small fraction of $\|\Delta V\|_F$, reducing the effective sensitivity.

For $V, V' \in \mathbb{R}^{d\times n}$ write $\Delta V := V - V'$. For $\varepsilon > 0$ and $\mu \ge 0$, define the Gaussian two-sided tail function

$$T(\varepsilon;\mu) := \Phi\left(\frac{-\varepsilon+\mu/2}{\sqrt{\mu}}\right) + 1 - \Phi\left(\frac{\varepsilon+\mu/2}{\sqrt{\mu}}\right),$$

where Φ is the standard normal CDF and $T(\varepsilon;0) = 0$.

Theorem 5 (Privacy of M2 in the small r-regime)**.** *Fix $V, V' \in \mathbb{R}^{d\times n}$ and write $s = \min(n,d)$. Then for any $\alpha \in (0,1)$ and any $\varepsilon > 0$, $0 < r < d$, the mechanism $\mathcal{A}_2$ is $(\varepsilon, \delta_\alpha(\varepsilon))$-DP with*

$$\delta_\alpha(\varepsilon) \le T\left(\varepsilon; \frac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}\right) + s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right],$$

where $I_\alpha(a,b)$ is the regularized incomplete Beta function. Moreover, the same privacy guarantee holds for $M'(V+\Xi)$, where M' is obtained by orthonormalizing M.

4.2.1 PROOF SKETCH

Fix V, V' and condition on M. Let $P_M = M(M^2)^\dagger M$ denote the orthogonal projector onto $\mathrm{col}(M)$, and let P_M and Q_M be the conditional output laws under inputs V and V', respectively. Conditional on M, both outputs are Gaussian matrices supported on $\mathrm{col}(M)$, with the same covariance and mean difference $M\Delta V$. Thus the privacy loss depends on ΔV only through its component in $\mathrm{col}(M)$. More precisely, for $\ell_M(Y) := \log \frac{d\mathsf{P}_M}{d\mathsf{Q}_M}(Y),$ with $Y \sim \mathsf{P}_M$, Lemma 17 gives

$$\ell_M(Y) \mid M \sim \mathcal{N}\left(\frac{\rho(M)}{2}, \rho(M)\right), \qquad \rho(M) := \frac{\|P_M \Delta V\|_F^2}{\sigma_G^2}.$$

The same two-sided privacy-loss tail is obtained when the output is drawn from Q_M. Therefore, for the good-output event $\mathcal{Y}_\varepsilon(M) := \{y : |\ell_M(y)| \le \varepsilon\}$, the conditional failure probability is exactly $T(\varepsilon; \rho(M))$. Now define the good-projection event

$$\mathcal{G}_\alpha := \left\{M : \|P_M \Delta V\|_F^2 \le \alpha \|\Delta V\|_F^2\right\}.$$

On $\mathcal{G}_\alpha$, $\rho(M) \le \frac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}$. Since $T(\varepsilon; \rho)$ is nondecreasing in ρ, conditioning on any $M \in \mathcal{G}_\alpha$ gives

$$\Pr\big(|\ell_M(Y)| > \varepsilon \mid M \in \mathcal{G}_\alpha\big) \le T\left(\varepsilon; \frac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}\right).$$

Averaging over M, the total privacy error is therefore bounded by this Gaussian tail plus the failure probability $\Pr(\mathcal{G}_\alpha^c)$. It remains to control $\Pr(\mathcal{G}_\alpha^c)$. Let $s = \mathrm{rank}(\Delta V) \le \min\{n, d\}$, and write $\Delta V = \sum_{\ell=1}^s \tau_\ell u_\ell w_\ell^\top$. Then $\|P_M \Delta V\|_F^2 = \sum_{\ell=1}^s \tau_\ell^2 \|P_M u_\ell\|_2^2$. Hence $\mathcal{G}_\alpha$ holds whenever $\|P_M u_\ell\|_2^2 \le \alpha$ for all $\ell \in [s]$. Since $\mathrm{col}(M)$ is a Haar-random rank-r subspace, for each fixed unit vector u_ℓ, $\|P_M u_\ell\|_2^2 \sim \mathrm{Beta}\left(\frac{r}{2}, \frac{d-r}{2}\right)$. A union bound gives

$$\Pr(\mathcal{G}_\alpha^c) \le h\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right] \le s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right].$$

Combining the two bounds yields Theorem 5.

4.2.2 COMPARISON WITH BASELINES

The classical Gaussian mechanism ignores the randomness in M and uses the full sensitivity $\|\Delta V\|_F$, giving

$$\delta_{\mathrm{Gauss}}(\varepsilon) = T\left(\varepsilon; \frac{\|\Delta V\|_F^2}{\sigma_G^2}\right).$$

Our bound replaces this by a trade-off: on $\mathcal{G}_\alpha$, the effective squared sensitivity is only $\alpha\|\Delta V\|_F^2$, while $\mathbb{P}(\mathcal{G}_\alpha^c)$ contributes the residual term. Optimizing over α yields the gains shown in Figure 3; in the small-r regime, a random rank-r subspace captures only about r/d of the energy of a fixed direction. Corollary 1 makes this improvement precise. Moreover, for the orthonormalized mechanism $P_M(V + \Xi)$, where P_M is the orthogonal projector onto $\mathrm{col}(M)$, $\|P_M\|_{\mathrm{op}} = 1$, so this reduction in additive noise is not subsequently amplified by the projection.

To relate this to common training procedures, note that the Gaussian baseline above is the privacy accounting method used in DP-SGD with $V = \nabla_W \mathcal{L}(W_t)^\top$. Write E as standard Gaussian noise matrix, then the gradients used in one update step in DP-SGD and with our M2 algorithm with orthonormalized M are appproximately,

$$V + \sigma_G E \qquad \text{and} \qquad P_M\Big(V + \sqrt{r/d}\,\sigma_G E\Big),$$

respectively, ignoring the low-probability projection-failure event. Hence, the orthonormalized mechanism incurs an approximation error of the full gradient $V - P_M V$, while requiring substantially less additive noise. Whether this improves utility depends on whether the reduction in noise outweighs the projection error $\|(I - P_M)V\|_F$.

Hence, our privacy improvement does not imply uniformly better utility, since the optimization trajectory also depends on r, clipping, and the geometry of the gradients. When the sampled subspace preserves the important gradient directions, the reduced noise can yield an advantage over DP-SGD.

Conversely, when relevant gradient directions vary over training or form a non-negligible spectral tail, repeatedly resampling the projection can cover a broader set of directions over time than a fixed-subspace method such as DP-LoRA-FA. Section G.1 discuss the difference between our Noisy Wishart mechanism M2 and DP-LoRA-FA, while Section H provides additional experimental results to illustrate the above effects on synthetic and real datasets. A full characterization of the resulting privacy-utility trade-off of the noisy Wishart mechanism is left for future work.

5 Discussion and Open Questions

Related works Random projections are widely exploited in the privacy literature. Some works, such as Kenthapadi et al. (2013); Li & Li (2023), explore the privacy of JL-style projections or random sign flipping. However, these approaches typically do not treat the projection's randomness as part of the privacy mechanism: they publish the projection matrix and regard its randomness as public information. Other lines of work use projections mainly for dimensionality reduction to improve the privacy–utility trade-off by reducing the dimension dependence in convergence guarantees (Jiang et al., 2025; Kasiviswanathan, 2021; Li et al., 2011; Sheffet, 2019). The intuition is similar to our Theorem 5, however, those results improve utility in expectation or with high probability over algorithmic randomness, whereas we directly establish a *privacy amplification* at fixed noise and characterize its dependence on the rank r.

Far fewer works exploit projection randomness itself as a privacy resource. Lev et al. (2025) show that Gaussian sketching can amplify privacy: releasing a Gaussian sketch with additive noise ($Z^\top V + \Xi$) yields improved privacy guarantees under a "scale" assumption on the singular values of the data matrix. Our setting is LoRA-motivated and geometrically different: we project with a Wishart matrix $M = ZZ^\top$ and release $MV + \Xi$ (M1). As M is PSD, the induced geometry differs from Gaussian sketching, leading to a different analysis and assumptions. Nevertheless, the takeaway is the same: intrinsic projection randomness plus modest noise can yield stronger privacy guarantees than additive noise alone.

There is also a line of work on privacy amplification via compression. In particular, Jin & Dai (2025) shows that by compressing the gradient to their signs, amplifies privacy guarantees. Perhaps most related to our work, Hao et al. (2024) interpret LoRA as gradient compression via (re-sampled) low-dimensional random projections, but their motivation is primarily *efficiency*, not privacy.

A recent preprint argues that LoRA can provide differential privacy by design through Gaussian sketching (Malekmohammadi & Farnadi, 2024). However, its post-processing argument does not apply to finite-dimensional matrix-valued updates.[3] In contrast, we analyze the released update directly: we give formal differential privacy guarantees for vector-valued updates, prove a sharp non-privacy result for noise-free matrix-valued updates, and show that, after adding small amounts of additive noise to the gradients, random projections yield *privacy amplification* beyond noise calibration alone.

Discussion The negative result for matrix-valued updates shows that noise-free matrix-valued LoRA-FA updates are not differentially private. This worst-case statement does not preclude weaker privacy notions. Apart from this, in Table 1 we observe perfect success for MIA when $r \geq 128$, but slightly lower success when r is extremely small. We attribute this gap primarily to a utility effect: very small-r LoRA does not achieve comparable test accuracy, making loss-based MIA less sensitive. Understanding LoRA's privacy risk under weaker adversaries, and disentangling genuine privacy gains from ineffective MIA adversaries at small r, is an interesting direction for future work.

On the positive side, we show that once additive noise is introduced, the intrinsic randomness of the projection mechanism can be leveraged to amplify privacy. However, for the large r case (Section 4.1) these alignment-based guarantees rely on data-dependent properties that may be difficult to verify or enforce in practice. An interesting direction for future work is to design algorithms or training objectives that promote such alignment, or to exploit it when present to improve privacy–accuracy trade-offs. Interestingly, our proof for large and small r rely on very different techniques; it is an interesting open question to study whether they can be combined to provide better overall guarantee.

[3] If the same random sketching matrix is used both to form the Gaussian sketch and to reconstruct the released update, then the reconstruction is not a post-processing of the sketch alone; the final matrix-valued update must be analyzed as part of the mechanism.

AI USE STATEMENT

The proofs of the positive privacy result for vector-valued outputs, the negative results for matrix-valued outputs, and the privacy guarantee for mechanism (M2) were developed by the authors. For mechanism (M1), the authors developed the key ideas of establishing privacy under one-column adjacency while accounting for the information about the shared projection revealed by the unchanged-column releases, and extending this guarantee to matrix adjacency through group privacy. We used LLMs to assist in developing the detailed proofs based on these ideas, and subsequently reviewed and verified the resulting arguments.

We also used generative AI to assist in writing experimental code and plotting, which was reviewed and verified by the authors. The authors take full responsibility for the content of the paper, including all theoretical results and experimental findings.

REFERENCES

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

Jeremiah Blocki, Avrim Blum, Anupam Datta, and Or Sheffet. The Johnson-Lindenstrauss transform itself preserves differential privacy. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pp. 410–419. IEEE, 2012.

Sanjoy Dasgupta and Anupam Gupta. An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, 2006.

Yongchang Hao, Yanshuai Cao, and Lili Mou. Flora: Low-rank adapters are secretly gradient compressors. In *ICML*, 2024.

Robert V. Hogg, Joseph W. McKean, and Allen T. Craig. *Introduction to Mathematical Statistics*. Pearson, 8 edition, 2019.

Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. Evaluating memorization in parameter-efficient fine-tuning. In *The Impact of Memorization on Trustworthy Foundation Models: ICML 2025 Workshop*, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

Zhanhong Jiang, Zahid Hasan, Nastaran Saadati, Balu, and Liu. Balancing utility and privacy: Dynamically private SGD with random projection. *Submitted to Transactions on Machine Learning Research*, 2025. Under review.

Richeng Jin and Huaiyu Dai. Noisy SIGNSGD is more differentially private than you (might) think. In *Forty-second International Conference on Machine Learning*, 2025.

Shiva Prasad Kasiviswanathan. Sgd with low-dimensional gradients with applications to private and distributed learning. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, pp. 1905–1915. PMLR, 2021.

Krishnaram Kenthapadi, Aleksandra Korolova, Ilya Mironov, and Nina Mishra. Privacy via the Johnson- Lindenstrauss transform. *Journal of Privacy and Confidentiality*, 5, Aug. 2013.

Omri Lev, Vishwak Srinivasan, Moshe Shenfeld, Katrina Ligett, Ayush Sekhari, and Ashia C. Wilson. The gaussian mixing mechanism: Renyi differential privacy via gaussian sketches. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Ping Li and Xiaoyun Li. Smooth flipping probability for differential private sign random projection methods. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yang D Li, Zhenjie Zhang, Marianne Winslett, and Yin Yang. Compressive mechanism: Utilizing sparse representation in differential privacy. In *Proceedings of the 10th annual ACM workshop on Privacy in the electronic society*, pp. 177–182, 2011.

Saber Malekmohammadi and Golnoosh Farnadi. Low-rank adaptation secretly imitates differentially private sgd. *arXiv preprint arXiv:2409.17538*, 2024.

Ilya Mironov. Rényi Differential Privacy . In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 2017.

Milad Nasr, Shuang Songi, Abhradeep Thakurta, Nicolas Papernot, and Nicholas Carlin. Adversary instantiation: Lower bounds for differentially private machine learning. In *2021 IEEE Symposium on Security and Privacy (SP)*, 2021.

Francesco Pinto, Yaxi Hu, Fanny Yang, and Amartya Sanyal. Pillar: How to make semi-private learning more effective. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2024.

Or Sheffet. Old techniques in differentially private linear regression. In Aurélien Garivier and Satyen Kale (eds.), *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, Proceedings of Machine Learning Research. PMLR, 2019.

Youbang Sun, Zitao Li, Yaliang Li, and Bolin Ding. Improving lora in privacy-preserving federated learning. In *International Conference on Learning Representations*, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Lele Zhang, Xiang Wang, Tao Zhang, Yang Cao, Ke Cheng, and Yulong Shen. Differentially Private Subspace Fine-Tuning for Large Language Models. In *arXiv:2601.11113*, 2026.

Longteng Zhang, Lin Zhang, Shaohuai Shi, Xiaowen Chu, and Bo Li. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303*, 2023.

A Appendix

B Mathematical Preliminaries

We recall several standard distributional facts, algebraic tools, and formal definitions that will be used throughout the proofs.

Lemma 3 (Post-processing property of DP)**.** *If $\mathcal{A}$ is (ε, δ)-DP and g is any (possibly randomized) mapping independent of the input dataset conditional on $\mathcal{A}(S)$, then the composition $g \circ \mathcal{A}$ is also (ε, δ)-DP.*

Lemma 4 (Basic composition)**.** *If $\mathcal{A}_1, \ldots, \mathcal{A}_K$ are each (ε, δ)-DP on the same domain and are run on the same dataset, then the tuple $(\mathcal{A}_1, \ldots, \mathcal{A}_K)$ is $(K\varepsilon, K\delta)$-DP.*

Definition 5 (Chi-square distribution.)**.** *A random variable V has a chi-square distribution with ν degrees of freedom, written $V \sim \chi^2_\nu$, if it can be represented as*

$$V = \sum_{i=1}^{\nu} Z_i^2, \qquad Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1).$$

Definition 6 (Student-t distribution as a ratio, Def. 8.3.1 in (Hogg et al., 2019))**.** *Let $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ be independent. Then the random variable*

$$T_\nu := \frac{Z}{\sqrt{V/\nu}}$$

follows a (central) Student-t distribution with ν degrees of freedom, denoted $T_\nu \sim t_\nu$. Equivalently,

$$\frac{Z}{\sqrt{V}} \overset{d}{=} \frac{1}{\sqrt{\nu}} T_\nu.$$

Lemma 5 (Corollary 7.3.2 in Vershynin (2018))**.** *Let A be an $m \times n$ matrix with independent $N(0,1)$ entries. Then, for $t \geq 0$, we have*

$$\mathbb{P}\left[\|A\| \geq \sqrt{m} + \sqrt{n} + t\right] \leq 2e^{-ct^2}.$$

Lemma 6. *Let $A \in \mathbb{R}^{m \times r}$ has i.i.d. $N(0,1)$ entries with $m > r$ and let $\sigma_1(A) \geq ... \geq \sigma_r(A)$ be its singular value, then for any $t \geq 0$,*

$$\mathbb{P}\left[\sigma_r(A) \leq \sqrt{m} - \sqrt{r} - t\right] \leq e^{-\frac{t^2}{2}}$$

Lemma 7 (Spectrum of a rank-r Wishart matrix, Thm. 4.6.1 Vershynin (2018))**.** *Let $Z \in \mathbb{R}^{d \times r}$ have i.i.d. $\mathcal{N}(0,1)$ entries and assume $d \geq r$. Define the (normalized) Wishart matrix*

$$W \;:=\; \frac{1}{r} Z Z^\top \in \mathbb{R}^{d \times d}.$$

Then $\operatorname{rank}(W) = r$, so W has exactly $d - r$ zero eigenvalues. Moreover, for every $0 \leq t \leq \sqrt{d} - \sqrt{r}$, with probability at least $1 - 2e^{-t^2/2}$,

$$\left(\sqrt{\frac{d}{r}} - 1 - \frac{t}{\sqrt{r}}\right)^2 \;\leq\; \lambda^+_{\min}(W) \;\leq\; \lambda_{\max}(W) \;\leq\; \left(\sqrt{\frac{d}{r}} + 1 + \frac{t}{\sqrt{r}}\right)^2, \tag{5}$$

where $\lambda^+_{\min}(W)$ denotes the smallest nonzero *eigenvalue of W. Equivalently, the nonzero spectrum of W lies in the interval above, i.e.*

$$\operatorname{spec}(W) \setminus \{0\} \subseteq \left[\left(\sqrt{\frac{d}{r}} - 1 - \frac{t}{\sqrt{r}}\right)^2, \left(\sqrt{\frac{d}{r}} + 1 + \frac{t}{\sqrt{r}}\right)^2\right]$$

with probability at least $1 - 2e^{-t^2/2}$.

Proof. By Vershynin's Gaussian singular value bound, for every $t \geq 0$, with probability at least $1 - 2e^{-t^2/2}$,

$$\sqrt{d} - \sqrt{r} - t \;\leq\; s_{\min}(Z) \;\leq\; s_{\max}(Z) \;\leq\; \sqrt{d} + \sqrt{r} + t.$$

Since the nonzero eigenvalues of $ZZ^\top$ equal $s_i(Z)^2$, the nonzero eigenvalues of $W = \frac{1}{r}ZZ^\top$ equal $\frac{1}{r}s_i(Z)^2$, giving equation 5. Finally, $\operatorname{rank}(ZZ^\top) = \operatorname{rank}(Z) = r$ almost surely, so W has $d - r$ zero eigenvalues. □

Proof. Use the Kronecker-vec identity

$$\operatorname{vec}(AX) = (I_k \otimes A)\operatorname{vec}(X),$$

(which is the special case of $\operatorname{vec}(BXC) = (C^\top \otimes B)\operatorname{vec}(X)$ with $B = A$ and $C = I_k$).

Therefore,

$$\operatorname{vec}(X)^\top (I_k \otimes A)\operatorname{vec}(X) = \operatorname{vec}(X)^\top \operatorname{vec}(AX).$$

Now apply the Frobenius inner-product identity

$$\operatorname{vec}(U)^\top \operatorname{vec}(V) = \operatorname{tr}(U^\top V),$$

with $U = X$ and $V = AX$. This gives

$$\operatorname{vec}(X)^\top \operatorname{vec}(AX) = \operatorname{tr}\big(X^\top (AX)\big) = \operatorname{tr}(X^\top AX).$$

□

Lemma 8 (Tail bound for random capture fraction)**.** *Let $r \leq d/2$ and let $B \sim \operatorname{Beta}\big(\frac{r}{2}, \frac{d-r}{2}\big)$. Fix any $\eta \in (0,1)$ and set $\alpha \coloneqq \frac{r(1+\eta)}{d} \in (0,1)$. Then*

$$\Pr(B > \alpha) \;\leq\; 2\exp\left(-\frac{\eta^2 r}{72}\right)$$

Proof. Let $X \sim \chi^2_r$ and $Y \sim \chi^2_{d-r}$ be independent. It is standard that $B \overset{d}{=} \frac{X}{X+Y}$. We claim that the event $\{B > \alpha\}$ is contained in the union of two simpler deviations:

$$\{B > \alpha\} \;\subseteq\; \Big\{X > (1+\eta/3)r\Big\} \;\cup\; \Big\{Y < (1-\eta/3)(d-r)\Big\}. \tag{6}$$

Indeed, suppose that both complementary events hold, i.e.

$$X \leq (1+\eta/3)r \qquad \text{and} \qquad Y \geq (1-\eta/3)(d-r).$$

Then

$$B = \frac{X}{X+Y} \;\leq\; \frac{(1+\eta/3)r}{(1+\eta/3)r + (1-\eta/3)(d-r)} = \frac{(1+\eta/3)r}{d - \eta/3(d-2r)}.$$

We claim that

$$\frac{(1+\eta/3)\,r}{d - \frac{\eta}{3}(d-2r)} \;\leq\; \frac{(1+\eta)\,r}{d} = \alpha.$$

Since $r > 0$ and $d - \frac{\eta}{3}(d-2r) > 0$, this is equivalent to

$$(1+\eta)\Big(d - \frac{\eta}{3}(d-2r)\Big) - \Big(1 + \frac{\eta}{3}\Big)\,d \;\geq\; 0.$$

Expanding, we obtain

$$(1+\eta)\Big(d - \frac{\eta}{3}(d-2r)\Big) - \Big(1 + \frac{\eta}{3}\Big)\,d = \frac{\eta(1-\eta)}{3}\,d \;+\; \frac{2\eta(1+\eta)}{3}\,r.$$

This is nonnegative since $r, d > 0$ and $\eta \in (0,1)$, which proves equation 6.

Next, we apply standard chi-square concentration: for $Z \sim \chi^2_k$ and any $t \in (0,1)$,

$$\Pr\big(Z > (1+t)k\big) \leq \exp\left(-\frac{t^2 k}{8}\right), \qquad \Pr\big(Z < (1-t)k\big) \leq \exp\left(-\frac{t^2 k}{8}\right). \tag{7}$$

Using equation 6 with $t = \eta/2$ and a union bound gives

$$\Pr(B > \alpha) \;\le\; \Pr\big(X > (1+\eta/3)r\big) + \Pr\big(Y < (1-\eta/3)(d-r)\big).$$

Applying equation 7 yields

$$\Pr\big(X > (1+\eta/3)r\big) \le \exp\!\left(-\frac{(\eta/3)^2 r}{8}\right) = \exp\!\left(-\frac{\eta^2 r}{72}\right),$$

and similarly

$$\Pr\big(Y < (1-\eta/2)(d-r)\big) \le \exp\!\left(-\frac{(\eta/3)^2 (d-r)}{8}\right) \le \exp\!\left(-\frac{\eta^2 r}{72}\right),$$

where the last inequality uses $d - r \ge r$ since $r \le d/2$. Therefore,

$$\Pr(B > \alpha) \le 2\exp\!\left(-\frac{\eta^2 r}{72}\right)$$

where the final step simply loosens the constant for a cleaner expression. □

Lemma 9. *For $x \ge 0$, $e^x - e^{-x} \ge 2x$.*

Proof. Define $f(x) := e^x - e^{-x} - 2x$. Then $f(0) = 0$, and

$$f'(x) = e^x + e^{-x} - 2 = \left(e^{x/2} - e^{-x/2}\right)^2 \ge 0.$$

Thus f is nondecreasing, so $f(x) \ge f(0) = 0$ for every $x \ge 0$. Therefore,

$$e^x - e^{-x} \ge 2x, \qquad x \ge 0.$$

□

Definition 7 (Orthogonal group)**.** *The orthogonal group is*

$$O(d) \;\coloneqq\; \{U \in \mathbb{R}^{d\times d} :\; U^\top U = I_d\}.$$

Definition 8 (Grassmannian)**.** *The Grassmannian $\mathrm{Gr}(d, r)$ is the set of all r-dimensional linear subspaces of $\mathbb{R}^d$:*

$$\mathrm{Gr}(d, r) \;\coloneqq\; \{S \subseteq \mathbb{R}^d :\; S \text{ is a linear subspace and } \dim(S) = r\}.$$

Definition 9 (Stiefel manifold)**.** *The (real) Stiefel manifold $V_{d,r}$ is the set of all $d \times r$ matrices with orthonormal columns:*

$$V_{d,r} \;\coloneqq\; \{Q \in \mathbb{R}^{d\times r} :\; Q^\top Q = I_r\}.$$

Definition 10 (Haar-uniformity on the Stiefel manifold)**.** *A random element $Q \in V_{d,r}$ is* Haar-uniform *on the Stiefel manifold if, for every $U \in O(d)$,*

$$UQ \;\overset{d}{=}\; Q.$$

Equivalently, the law of Q is the unique probability measure on $V_{d,r}$ that is invariant under the left action $Q \mapsto UQ$.

Definition 11 (Uniformity on the Grassmannian)**.** *A random subspace $S \in \mathrm{Gr}(d, r)$ is* uniform on the Grassmannian *if, for every $U \in O(d)$,*

$$US \;\overset{d}{=}\; S,$$

where $US \coloneqq \{Ux :\; x \in S\}$. Equivalently, the law of S is the unique probability measure on $\mathrm{Gr}(d, r)$ that is invariant under the action $S \mapsto US$.

C Privacy Analysis of Vector Projection Mechanism

Theorem 1. *For a dataset collection $\mathcal{D}$ and a query function f with outputs in $\mathbb{R}^d$ and $\|f(S)\|_2 = 1$ for all $S \in \mathcal{D}$, let $\rho > 0$ be the minimum alignment defined in Equation (2). Let $\eta \in (0, 1/3)$, $\nu = d + r - 1$, and*

$$K = \sqrt{\frac{1-\rho^2}{r}} t_r(1-\eta), \quad \Delta = \max\{\nu - \kappa_\nu(\eta/2), \kappa_\nu(1-\eta/2) - \nu\}.$$

If $\rho > K$, then the projection mechanism (Definition 1) with rank r is $(\varepsilon, 3\eta)$-DP on $\mathcal{D}$ for

$$\varepsilon \geq \frac{\nu + \Delta}{2}\left(\frac{1}{\rho - K} + \log(\rho - K) - 1\right) - (r - 1 + \Delta/2)\log(\rho - K).$$

Proof. Fix neighboring datasets $S \sim_H S'$ in $\mathcal{D}$ and set $v := (f(S))^\top, v' := (f(S'))^\top \in$, with $\|v\|_2 = \|v'\|_2 = 1$ and $\langle v, v' \rangle =: \rho \in (0, 1]$. Let $Z \in \mathbb{R}^{d \times r}$ have i.i.d. columns $z_k \sim \mathbb{N}(0, \frac{1}{r} I_d)$, and define $M := ZZ^\top$. The mechanism outputs $Y := f(S)M = v^\top M$. We will show that the mechanism $v \mapsto Mv$ is $(\varepsilon_\rho, \delta_\rho)$-DP. By post-processing property of DP (Lemma 3), Definition 1 ($V^\top \mapsto V^\top M$, equivalently, Definition 1) is also DP.

Let P and Q denote the laws of Y under inputs v and v' respectively (i.e.$Y = Mv$ and $Y = Mv'$), with densities p and q. By PDF of Mv (Lemma 11), $p(y)$ has the form

$$p(y) = C_{r,d,\sigma}(v^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2 v^\top y}\right)$$

on the half-space $\mathcal{Z}_p = \{Y : v^\top y > 0\}$. Here $p(\mathcal{Z}_p) = 1$. We can define the $\mathcal{Z}_q = \{Y : v'^\top y > 0\}$ similarly.

Then, we rewrite the privacy loss random variable as

$$\begin{aligned} L(y) = \log\frac{p(y)}{q(y)} &= \frac{d-r+1}{2}\log\left(\frac{(v')^\top y}{v^\top y}\right) + \frac{r\|y\|^2}{2v^\top y}\left(\frac{v^\top y}{(v')^\top y} - 1\right) \\ &= \frac{d-r+1}{2}\log A + \frac{B}{2}\left(\frac{1}{A} - 1\right) \end{aligned} \tag{8}$$

where $A := \frac{(v')^\top y}{v^\top y}$ and $B := \frac{r\|y\|^2}{v^\top y}$, with $y \sim p$. A quantifies relative alignment of y with v' vs v, and B quantifies how large y is compared to its v-projection. This means $L(y)$ will blow up when either B is very large or A is close to 0. Write $\xi = \langle v, v' \rangle \geq \rho$ for the fixed pair v, v'. Then, by Lemma 10 we have

$$A \overset{d}{=} \xi + \sqrt{\frac{1-\xi^2}{r}}\, T_r, \qquad B \overset{d}{=} \chi^2_{d+r-1}.$$

Then, on the event $|T_r| \leq t_r(1-\eta)$ whose probability is at least $1 - 2\eta$,

$$|A - 1| \leq 1 - \xi + \sqrt{\frac{1-\xi^2}{r}} t_r(1-\eta) \leq 1 - \rho + K. \tag{9}$$

Since $0 < \rho - K \leq 1$, this implies

$$\rho - K \leq A \leq 2 - \rho + K \leq \frac{1}{\rho - K}, \tag{10}$$

and

$$|\log A| \leq -\log(\rho - K). \tag{11}$$

By the definition of quantile function, and recall that $\Delta = \max\{\nu - k_\nu(\eta/2), k(1-\eta/2) - \nu\}$ with $\nu = d + r - 1$, we have

$$\mathbb{P}(|B - \nu| > \Delta) \leq \eta.$$

Define the good set $\mathcal{G} = \left\{ y : v^\top y > 0, \rho - K \le A \le \frac{1}{\rho - K}, |B - v| \le \Delta \right\}$, then

$$p(\mathcal{G}^c) \le 3\eta.$$

We note that any point in the good set $\mathcal{G}$ is also in the support of Q, so there is no need to separately account for the support mismatch between P and Q. By Equation (10), for any $y \in \mathcal{G}$, the following holds $y \in \mathcal{Z}_q$, as:

$$\frac{v'^\top y}{v^\top y} \ge \rho - K > 0, \quad v'^\top y \ge (\rho - K)(v^\top y) > 0.$$

This implies $\mathcal{G} \subset \mathcal{Z}_q$.

Next, we prove an upper bound on $L(y)$ for $y \in \mathcal{G}$. Write $t = \log A$,

$$\begin{aligned}
L(y) &= -(r-1)t - \frac{B-\nu}{2}t + \frac{B}{2}(e^{-t} - 1 + t) \\
&\le (r-1)\,|t| + \frac{B-\nu}{2}\,|t| + \frac{B}{2}(e^{|t|} - 1 + |t|) \\
&\overset{(a)}{\le} (r-1)\,|t| + \frac{B-\nu}{2}\,|t| + \frac{B}{2}(e^{|t|} - 1 - |t|) \\
&\overset{(b)}{\le} -\left(r - 1 - \frac{\Delta}{2}\right)\log(\rho - K) + \frac{\Delta + \nu}{2}(e^{|t|} - 1 - |t|) \\
&\overset{(c)}{\le} -\left(r - 1 - \frac{\Delta}{2}\right)\log(\rho - K) + \frac{\Delta + \nu}{2}\left(\frac{1}{\rho - K} + \log(\rho - K) - 1\right)
\end{aligned} \tag{12}$$

where step (a) follows by Lemma 9, and step (b) follows by Equation (11) and $|B - \nu| \le \Delta$. Finally, step (c) follows by the fact that $e^x - 1 - x$ is monotonically increasing for $x > 0$. This established that $L(y) \le \varepsilon$ for $y \in \mathcal{G}$.

Next, we convert this high probability upper bound on the privacy loss random variable to an approximate DP guarantee. For any measurable $\mathcal{Y}$,

$$p(\mathcal{Y}) \le e^\varepsilon q(\mathcal{Y} \cap \mathcal{G}) + p(\mathcal{G}^c) \le e^\varepsilon q(\mathcal{Y}) + 3\eta.$$

Swapping v, v' gives the same bound, which concludes the proof.

□

Lemma 10 (Distributional representation of A and B)**.** *Let $Z \in \mathbb{R}^{d \times r}$ have i.i.d. $N(0,1)$ entries and set $M := \frac{1}{r} ZZ^\top$. Fix unit vectors $v, v' \in \mathbb{R}^d$ with inner product $\rho := \langle v, v' \rangle \in [-1, 1]$, and let $y := Mv$. Then,*

$$A := \frac{(v')^\top y}{v^\top y} \overset{d}{=} \rho + \sqrt{\frac{1-\rho^2}{r}}\, T_r, \qquad B := \frac{r\|y\|^2}{v^\top y} \overset{d}{=} \chi^2_{d+r-1}.$$

where T_r denotes the Student-t distribution with r degree of freedom and χ^2_r denotes the chi-square distribution with r degree of freedom.

Proof. We first show that there exist independent random variables $K_1 \sim \chi^2_r$, $K_2 \sim N(0,1)$, $K_3 \sim \chi^2_{d-2}$ such that, jointly,

$$(A, B) \overset{d}{=} \left(\rho + \sqrt{1-\rho^2}\,\frac{K_2}{\sqrt{K_1}},\ K_1 + K_2^2 + K_3\right).$$

Recall that $M = ZZ^\top / r$ where Z is d by r with $i.i.d$ Gaussian entries $N(0,1)$. In particular, Let $Q \in \mathbb{R}^{d \times d}$ be an orthonormal matrix such that $Qv = e_1, Qv' = \rho e_1 + \sqrt{1-\rho^2} e_2$. Then, for $y \sim q$,

$$\begin{aligned}
A \coloneqq \frac{(v')^\top y}{v^\top y} &\overset{d}{=} \frac{(v')^\top ZZ^\top v}{v^\top ZZ^\top v} \overset{d}{=} \frac{(v')^\top QZZ^\top Q^\top v}{v^\top QZZ^\top Q^\top v} \\
&= \frac{(\rho e_1 + \sqrt{1-\rho^2} e_2)^\top ZZ^\top e_1}{e_1^\top ZZ^\top e_1} \\
&= \rho + \sqrt{1-\rho^2}\frac{e_2^\top ZZ^\top e_1}{e_1^\top ZZ^\top e_1} \overset{d}{=} \rho + \sqrt{1-\rho^2}\frac{g_2^\top g_1}{g_1^\top g_1}
\end{aligned}$$

where $g_1 = Ze_1, g_2 = Ze_2 \sim \mathcal{N}(0, I_r)$ and g_1 is independent of g_2.

$$\frac{g_2^\top g_1}{g_1^\top g_1} = \frac{g_2^\top \frac{g_1}{\|g_1\|}}{\|g_1\|} \overset{d}{=} \frac{K_2}{\sqrt{K_1}}$$

for $K_2 \sim N(0,1)$, $K_1 \sim \chi^2_r$ and K_2 independent of K_1. As $g_1 \sim \mathcal{N}(0, I_r)$, $\frac{g_1}{\|g_1\|}$ is independent of $\|g_1\|$. Thus, $g_2^\top g_1 / \|g_1\|$ is independent of $\|g_1\|$ and the last equality follows. So by Definition 6 we can write $A \overset{d}{=} \rho + \sqrt{\frac{1-\rho^2}{r}} T_r$ where T_r is a random variable following the student-t distribution with degree of freedom r.

$$\begin{aligned} B &:= \frac{r\|y\|^2}{2v^\top y} = \frac{e_1 ZZ^\top ZZ^\top e_1}{e_1 ZZ^\top e_1} \\ &= \frac{g_1^\top Z^\top Z g_1}{g_1^\top g_1} = \frac{\sum_{i=1}^d (g_i^\top g_1)^2}{g_1^\top g_1} \\ &= g_1^\top g_1 + \sum_{i=2}^d \left(g_i^\top \frac{g_1}{\|g_1\|} \right)^2 \overset{d}{=} \chi^2_{r+d-1} \end{aligned}$$

where the last inequality follows as conditional on g_1 we have that for $u = g_1/\|g_1\|$, $g_i^\top u \sim \mathcal{N}(0,1)$ and therefore conditional on g_1 for each $i \geq 2$ we have $\sum_{i=2}^d (g_i^\top u)^2 \sim \chi^2_{d-1}$. And lastly $\sum_{i=2}^d (g_i^\top u)^2 \sim \chi^2_{d-1}$ is independent of $g_1^\top g_1 = \|g_1\|_2^2$, so we can remove the conditioning. □

Lemma 11 (PDF of Mv)**.** *Let $z_1, ..., z_r$ be i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ with $r \geq 1$, $M = \sum_{i=1}^r z_i z_i^T$, then for $v \in \mathbb{R}^d$ with $\|v\| = 1$ and $y \in \mathbb{R}^d$ such that $v^\top y > 0$,*

$$\mathbb{P}(Mv = y) = C_{r,d,\sigma} (v^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2 v^\top y} \right)$$

where $C_{r,d,\sigma} = \frac{1}{2^{r/2}\Gamma(r/2)\sigma^{d+r-1}(2\pi)^{(d-1)/2}}$

Proof. For $Y = Mv = \sum_{i=1}^r z_i z_i^\top v$, let $a_i = z_i^\top v$, and $u_i = (I - vv^\top) z_i =: P_\perp z_i$.

Therefore, we can write $z_i = vv^\top z_i + (I - vv^\top) z_i = v a_i + u_i$, and Y as

$$Y = \sum_{i=1}^r z_i z_i^\top v = \sum_{i=1}^r (v a_i + u_i)(a_i v^\top + u_i^\top) v = \sum_{i=1}^r v a_i^2 + u_i a_i,$$

where $a_i \sim \mathcal{N}(0, \sigma^2), u_i \sim \mathcal{N}(0, \sigma^2 P_\perp)$. Further a_i and u_i are independent as

$$\mathrm{Cov}(a_i, u_i) = \mathbb{E}[a_i \cdot u_i] = \mathbb{E}[z_i^T v P_\perp z_i] = \mathbb{E}[P_\perp z_i z_i^T v] = P_\perp \mathbb{E}[z_i z_i^T] v = \sigma^2 P_\perp v = 0.$$

Let $S = \sum_{i=1}^r a_i^2$, then

$$S \sim \sigma^2 \chi^2_r, \quad Y|S = s \overset{d}{=} N\left(sv, \sigma^2 s P_\perp\right).$$

Define $U \in \mathcal{R}^{d \times (d-1)}$ as $[u_1 \cdots u_{d-1}] \in \mathbb{R}^{d \times (d-1)}$, where $\{u_1, \ldots, u_{d-1}\}$ is an orthonormal basis of $v^\perp$ then

$$U^\top U = I_{d-1}, \quad UU^\top = I - vv^\top = P_\perp.$$

Let $Y_\perp = U^\top Y$, then

$$S \sim \sigma^2 \chi^2_r, \quad Y_\perp | S = s \overset{d}{=} N\left(0, \sigma^2 s U^\top P_\perp U\right) \overset{d}{=} N\left(0, \sigma^2 s I_{d-1}\right).$$

We note the last inequality follows from substituting $P_\perp$ and noticing

$$U^\top v = (U^\top U) U^\top v = U^\top (UU^\top v) = U^\top 0 = 0.$$

Therefore, for $y \in \{y \in \mathbb{R}^d : y^\top v > 0\}$,

$$\mathbb{P}\left[S = s, Y_\perp = y_\perp\right] = \mathbb{P}\left[S = s\right] \mathbb{P}\left[Y_\perp = y_\perp | S = s\right] \tag{13}$$

$$f_{S\sigma^2}(s) = f_S\left(\frac{s}{\sigma^2}\right) = \frac{s^{r/2-1}e^{-\frac{s}{2\sigma^2}}}{2^{r/2}\Gamma(r/2)\sigma^{r-2}} \tag{14}$$

$$f_{Y_\perp|S}(y_\perp|S=s) = (2\pi)^{-\frac{d-1}{2}}\left(\sigma^2 s\right)^{-\frac{d-1}{2}}\exp\left(-\frac{\|y_\perp\|^2}{2\sigma^2 s}\right) \tag{15}$$

$$f_{S,Y_\perp}(s,y_\perp) = C_{r,d,\sigma}s^{\frac{r-d+1}{2}}e^{-\frac{s^2+\|y_\perp\|^2}{2\sigma^2 s}}, \quad C_{r,d,\sigma} = \frac{1}{2^{r/2}\Gamma(r/2)\sigma^{d+r-1}(2\pi)^{(d-1)/2}} \tag{16}$$

As we can write $\begin{pmatrix} S \\ Y_\perp \end{pmatrix} = \begin{pmatrix} v^\top \\ U^\top \end{pmatrix} Y$ (using point wise multiplication), let $Q = \begin{pmatrix} v^\top \\ U^\top \end{pmatrix}$. One can easily verify that $Q^\top = Q^{-1}$ and

$$Y = Q^\top \begin{pmatrix} S \\ Y_\perp \end{pmatrix}. \tag{17}$$

By changing the variables from $(S, Y_\perp)$ to Y with Equation (17), we get the probablity density function for Y when $y^\top v \geq 0$ ($s > 0$), i.e.

$$\mathbb{P}(Y=y) = C_{r,d,\sigma}(v^\top y)^{\frac{r-d+1}{2}}\exp\left(-\frac{r\left\|U^\top y\right\|^2}{2v^\top y}\right)$$

$$\exp\left(-\frac{r}{2}\left(s+\frac{\left\|U^\top y\right\|^2}{v^\top y}\right)\right) = \exp\left(-\frac{r}{2v^\top y}\left(y^\top(vv^\top+UU^\top)y\right)\right) = \exp\left(-\frac{r\|y\|^2}{2v^\top y}\right) \tag{18}$$

So we get

$$f_Y(y) = C_{r,d,\sigma}(v^\top y)^{\frac{r-d+1}{2}}\exp\left(-\frac{\|y\|^2}{2\sigma^2 v^\top y}\right) \tag{19}$$

□

C.1 Vector Privacy Amplification and Applications

Privacy amplification by increasing effective alignment The privacy guarantees for random projection in Theorem 1 can be strengthened by introducing a simple *pre-processing strategy*. We add uniform noise from a d-dimensional ball of radius $\gamma/2$ to $f(S)$ before applying the projection. Specifically,

$$M\left(f(S)+\frac{\gamma z}{\|z\|}\right), \quad z \sim N(0, \mathbf{I}_d)$$

This improves the effective alignment, especially when the original alignment ρ is small (or even negative) and in high-dimensional settings.

Lemma 12. *Let $v, v' \in \mathbb{R}^d$ be two unit vectors with with $\cos\angle(v,v') = v^\top v' \geq \rho$, $z \in \mathcal{N}(0, \mathbf{I}_d)$, $\delta > 0$ and $\gamma > \frac{1-\rho}{1+\rho}\sqrt{\frac{2}{d}\log\frac{8}{\delta}}$, then with probability at least $1-\delta$, we have*

$$\cos\left(\angle\left(v+\frac{\gamma z}{\|z\|_2}, v'+\frac{\gamma z}{\|z\|_2}\right)\right) \geq \rho + s > \rho,$$

where $s = \frac{(1-\rho)\gamma^2 - 4\gamma\sqrt{\frac{2}{d}\ln\frac{8}{\delta}}}{1+\gamma^2+2\gamma\sqrt{\frac{2}{d}\ln\frac{8}{\delta}}}$.

We observe that achieving a fixed target improvement s in alignment requires choosing a larger γ and adding more noise when the minimum alignment ρ is large (i.e., when the original vectors are already well aligned).

C.2 APPLICATIONS

In this section, we describe another possible applications of projected mechanism.

Projected gradient descent (RP–GD). Analogous to DP–GD, which privatises gradients by additive noise, we privatise the *average gradient direction* via the projection mechanism and then take a descent step with the projected output. Concretely, at each step, sample a fresh $M_t \sim \mathsf{W}_d\big(\sigma^2 I_d, r\big)$

$$w_{t+1} \;=\; w_t \;-\; \eta\, M_t \nabla \mathcal{L}(w_t)\,.$$

Private Retrival Aother possible application is to publish *private embeddings for retrieval tasks*. Given a unit-normalised average embedding v, sample $M \sim \mathsf{W}_d\left(\sigma^2 I_d, r\right)$ and release the $y = Mv$. The retrieval system maintains its catalogue $\{u_j\} \subset \mathbb{R}^d$ unchanged and ranks by standard dot products $\langle u_j, y\rangle = \langle u_j, Mv\rangle$. Since $\mathbb{E}[M] = r\sigma^2 I_d$ (unlike projections like the JL transformation) and $\|M - r\sigma^2 I_d\|$ concentrates for moderate r, these scores approximate a constant multiple of $\langle u_j, v\rangle$, preserving top-k ordering up to a small distortion that vanishes as r grows. This is useful for various modern retrival applications, where the embedding v is computed as an average of multiple embeddings. The same pattern applies to *releasing class/cohort prototypes*: compute the cohort mean, normalise and release $y = Mv$. In short, any application where the original embedding is an average embedding and final utility is measured with respect to cosine angle is a good fit for the projection mechanism.

D MATRIX PROJECTION MECHANISM IS NOT PRIVATE

In this section, we prove Lemma 13 that directly implies Proposition 2.

Lemma 13 (Failure of DP under matrix non-alignment)**.** *Let $d, m > 1$, $r \geq 1$, and let $Z, Z' \in \mathbb{R}^{d\times r}$ be independent random matrices with i.i.d. Gaussian entries. Consider the matrix-valued projection mechanism*

$$\mathcal{A}(S) = ZZ^\top f(S).$$

Suppose there exist neighboring datasets S, S' such that, writing $V = f(S), V' = f(S')$, one of the following conditions holds:

1. $\operatorname{rank}(V) < m$ *and there exists* $a \in \ker(V)$ *such that* $V'a \neq 0$*;*
2. $\operatorname{rank}(V) = m$ *and* $V' \neq cV$ *for every* $c \in \mathbb{R}$*.*

Then there exists a measurable event $E \subseteq \mathbb{R}^{d\times m}$ such that

$$\Pr_{Z}(ZZ^\top V \in E) = 1, \qquad \Pr_{Z'}(Z'Z'^\top V' \in E) = 0.$$

Consequently, $\mathcal{A}$ is not (ε, δ)-DP for any finite ε and any $\delta < 1$.

Proof. We prove the two cases separately. Throughout, Z and Z' denote independent copies of the projection randomness. Differential privacy compares the marginal laws of $ZZ^\top V$ and $Z'Z'^\top V'$, so no coupling between Z and Z' is required.

Case 1: $\operatorname{rank}(V) < m$**.** By assumption, there exists $a \in \ker(V)$ such that $V'a \neq 0$. Define the measurable event

$$E := \{Y \in \mathbb{R}^{d\times m} : Ya = 0\}.$$

For dataset S, every realization of the mechanism satisfies

$$(ZZ^\top V)a = ZZ^\top Va = 0,$$

because $Va = 0$. Hence

$$\Pr_{Z}(ZZ^\top V \in E) = 1.$$

For dataset S', let $b := V'a$. By assumption $b \neq 0$. Then

$$Z'Z'^\top V'a = 0 \iff Z'Z'^\top b = 0 \iff Z'^\top b = 0.$$

The last equivalence follows because

$$b^\top Z' Z'^\top b = \|Z'^\top b\|_2^2.$$

Since $b \neq 0$ and the columns of Z' are Gaussian with densities, each inner product $\langle z_i', b\rangle$ is a nondegenerate Gaussian random variable. Therefore

$$\Pr_{Z'}(Z'^\top b = 0) = 0.$$

Thus

$$\Pr_{Z'}(Z' Z'^\top V' \in E) = 0.$$

Case 2: $\operatorname{rank}(V) = m$**.** Define the event

$$E := \{Y \in \mathbb{R}^{d\times m} : V^\top Y = (V^\top Y)^\top\}.$$

For dataset S, every output $Y = ZZ^\top V$ satisfies

$$V^\top Y = V^\top ZZ^\top V,$$

which is symmetric. Hence

$$\Pr_{Z}(ZZ^\top V \in E) = 1.$$

We now show that

$$\Pr_{Z'}(Z' Z'^\top V' \in E) = 0.$$

The condition $Z' Z'^\top V' \in E$ is equivalent to

$$V^\top Z' Z'^\top V' = (V^\top Z' Z'^\top V')^\top,$$

or

$$V^\top Z' Z'^\top V' = V'^\top Z' Z'^\top V.$$

Equivalently, all entries of the skew-symmetric matrix

$$V^\top Z' Z'^\top V' - V'^\top Z' Z'^\top V$$

must vanish. These entries are polynomial functions of the entries of Z'.

It remains to check that at least one of these polynomials is not identically zero. Suppose, toward contradiction, that all of them were identically zero. Then for every $z \in \mathbb{R}^d$, taking a matrix Z' whose first column is z and whose remaining columns are zero would give

$$V^\top zz^\top V' = V'^\top zz^\top V.$$

Equivalently,

$$(V^\top z)(V'^\top z)^\top$$

is symmetric for every $z \in \mathbb{R}^d$.

Let $u := V^\top z$ and $w := V'^\top z$. Since V has full column rank, the map $z \mapsto V^\top z$ is surjective onto $\mathbb{R}^m$. The symmetry of $uw^\top$ implies that w is collinear with u whenever $u \neq 0$.

We first show that V' has no component orthogonal to the column space of V. Fix any $h \in \ker(V^\top)$. For any z with $V^\top z = u \neq 0$, also $V^\top(z+h) = u$. Thus both $V'^\top z$ and $V'^\top(z+h) = V'^\top z + V'^\top h$ must be collinear with u. Since u can be chosen arbitrarily in $\mathbb{R}^m$, this forces $V'^\top h = 0$ for every $h \in \ker(V^\top)$. Hence the columns of V' lie in the column space of V, so there exists a matrix $C \in \mathbb{R}^{m\times m}$ such that

$$V' = VC.$$

Now $V'^\top z = C^\top V^\top z = C^\top u$. The symmetry condition becomes

$$u(C^\top u)^\top \quad \text{is symmetric for every } u \in \mathbb{R}^m.$$

This means $C^\top u$ is collinear with u for every u. Since $m > 1$, the only linear maps with this property are scalar multiples of the identity. Thus $C = cI_m$ for some $c \in \mathbb{R}$, and therefore

$$V' = cV,$$

contradicting the assumption of Case 2.

Therefore at least one skew-symmetry polynomial is nonzero. Since Z' has a density with respect to Lebesgue measure and the zero set of a nonzero polynomial has Lebesgue measure zero, we conclude that

$$\Pr_{Z'}\big(V^\top Z' Z'^\top V' = V'^\top Z' Z'^\top V\big) = 0.$$

Hence

$$\Pr_{Z'}(Z' Z'^\top V' \in E) = 0.$$

In either case, we have found an event E such that

$$\Pr_{Z}(Z Z^\top V \in E) = 1, \qquad \Pr_{Z'}(Z' Z'^\top V' \in E) = 0.$$

If the mechanism were (ε, δ)-DP, then for this neighboring pair and this event E we would have

$$1 = \Pr_{Z}(Z Z^\top V \in E) \le e^{\varepsilon} \Pr_{Z'}(Z' Z'^\top V' \in E) + \delta = \delta.$$

This is impossible whenever $\delta < 1$. Therefore the mechanism is not (ε, δ)-DP for any finite ε and any $\delta < 1$. □

D.1 Privacy implication for standard LoRA

Here, we detail how Section 3.2 rules out intrinsic privacy for full LoRA. In the LoRA-FA setting, the noise-free effective weight update forms a matrix-valued Wishart projection:

$$W_{t+1} - W_t = -\eta \nabla_W L(W_t)(A^\top A) \tag{20}$$

where $A^\top A$ represents a rank-r Wishart draw. We observe that standard LoRA exhibits identical behavior at initialization. Under the standard initialization $B_0 = 0$, the first step of full LoRA (even with a trainable A) yields the update:

$$W_1 - W_0 = -\eta \nabla_W L(W_0) A_0^\top A_0 \tag{21}$$

Consequently, we can post-process the LoRA output W_1 by $(W_0 - W_1)\eta$ to get the output of projection mechanism. By the post-processing property of DP, if the noise-free LoRA mechanism were (ε, δ)-DP, then the projection mechanism would also be (ε, δ)-DP. However, this contradicts the results from Section 3.2, which establishes that such projections are not private. Therefore, we conclude that LoRA is not private without additive noise.

E Privacy Analysis of Matrix Projection Mechanism

E.1 Proofs in the Large r Regime

Notation and setup. In this section we denote neighboring datasets as $V \sim_A^{(j)} V'$ and let $S = \mathrm{span}(V_{-j})$ with $\dim(S) = s$. Let $U \in \mathbb{R}^{d\times s}$ have orthonormal columns spanning S, and let $U_\perp \in \mathbb{R}^{d\times(d-s)}$ have orthonormal columns spanning $S^\perp$.

Let $Z \in \mathbb{R}^{d\times r}$ have i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries. We define the Gaussian blocks

$$G := U^\top Z \in \mathbb{R}^{s\times r}, \qquad W := U_\perp^\top Z \in \mathbb{R}^{(d-s)\times r}.$$

By rotational invariance of the Gaussian and orthogonality of $[U\ U_\perp]$, the matrices G and W are independent and have i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries, and

$$Z = UG + U_\perp W.$$

Further, we define $H := \mathrm{rowspan}(G) \subseteq \mathbb{R}^r$, let P_H and $P_H^\perp$ be the orthogonal projectors onto H and $H^\perp$, and set

$$Z_\parallel := Z P_H, \qquad Z_\perp := Z P_H^\perp, \qquad M_\parallel := Z_\parallel Z_\parallel^\top, \qquad M_\perp := Z_\perp Z_\perp^\top,$$

with $p := \dim(H) = \mathrm{rank}(G) \le \min\{s, r\}$.

Lemma 14 (Exact orthogonal split)**.** *We have $M = M_{\|} + M_{\perp}$ and $Z_{\|} Z_{\perp}^{\top} = 0$. Moreover,*

$$U^{\top} Z_{\perp} = 0 \quad \textit{and hence} \quad \mathrm{range}(M_{\perp}) \subseteq S^{\perp}.$$

In particular, $M_{\perp} a = 0$ for all $a \in S$.

Proof. Recall that P_H and $P_H^{\perp}$ are orthogonal projectors onto H and $H^{\perp}$, hence

$$P_H^2 = P_H, \quad (P_H^{\perp})^2 = P_H^{\perp}, \quad P_H^{\top} = P_H, \quad (P_H^{\perp})^{\top} = P_H^{\perp}, \quad \text{and} \quad P_H + P_H^{\perp} = I_r, \; P_H P_H^{\perp} = 0.$$

By definition, $Z_{\|} = Z P_H$ and $Z_{\perp} = Z P_H^{\perp}$. Therefore,

$$M \;=\; Z Z^{\top} \;=\; Z(P_H + P_H^{\perp})(P_H + P_H^{\perp})^{\top} Z^{\top}.$$

Using symmetry of the projectors and expanding, we obtain

$$\begin{aligned} M &= Z(P_H + P_H^{\perp})(P_H + P_H^{\perp}) Z^{\top} \\ &= Z P_H Z^{\top} + Z P_H^{\perp} Z^{\top} + Z P_H P_H^{\perp} Z^{\top} + Z P_H^{\perp} P_H Z^{\top} \\ &= Z P_H Z^{\top} + Z P_H^{\perp} Z^{\top} \;=\; Z_{\|} Z_{\|}^{\top} + Z_{\perp} Z_{\perp}^{\top} \;=\; M_{\|} + M_{\perp}, \end{aligned}$$

since $P_H P_H^{\perp} = P_H^{\perp} P_H = 0$. This proves $M = M_{\|} + M_{\perp}$.

Next, the cross term vanishes:

$$Z_{\|} Z_{\perp}^{\top} \;=\; (Z P_H)(Z P_H^{\perp})^{\top} \;=\; Z P_H (P_H^{\perp})^{\top} Z^{\top} \;=\; Z P_H P_H^{\perp} Z^{\top} \;=\; 0.$$

We now show $U^{\top} Z_{\perp} = 0$. Using $G = U^{\top} Z$ and $Z_{\perp} = Z P_H^{\perp}$,

$$U^{\top} Z_{\perp} \;=\; U^{\top} Z P_H^{\perp} \;=\; G P_H^{\perp}.$$

By definition $H = \mathrm{rowspan}(G)$, hence every row of G lies in H. Projecting any vector in H onto $H^{\perp}$ yields zero, so $G P_H^{\perp} = 0$, and therefore $U^{\top} Z_{\perp} = 0$.

Finally, since $M_{\perp} = Z_{\perp} Z_{\perp}^{\top}$, we have

$$\mathrm{range}(M_{\perp}) \subseteq \mathrm{range}(Z_{\perp}).$$

Moreover, for any $x \in S$ we can write $x = U\alpha$ for some $\alpha \in \mathbb{R}^s$, and thus

$$Z_{\perp}^{\top} x \;=\; Z_{\perp}^{\top} U \alpha \;=\; (U^{\top} Z_{\perp})^{\top} \alpha \;=\; 0.$$

Hence

$$M_{\perp} x \;=\; Z_{\perp} Z_{\perp}^{\top} x \;=\; Z_{\perp}(Z_{\perp}^{\top} x) \;=\; 0,$$

which shows $M_{\perp} a = 0$ for all $a \in S$. Equivalently, $\mathrm{range}(M_{\perp}) \subseteq S^{\perp}$. □

We define the output variables of interest as

$$X \coloneqq M v_j + \xi_j \in \mathbb{R}^d, \qquad Y \coloneqq \left[\, M v_k + \xi_k \,\right]_{k \neq j} \in \mathbb{R}^{d \times (n-1)}.$$

where v_j denotes the jth column of $V \in \mathbb{R}^{d \times n}$ and $\{\xi_i\}_{i=1}^n$ are independent noise vectors.

Lemma 2 (Residual DP bound)**.** *Let $s \coloneqq \dim(S)$ and $q \coloneqq r - p$, and $\nu = d - s + q - 1$ where $p = \mathrm{rank}(G)$. Fix $\rho_{\perp} \in (0, 1]$, $\delta' \in (0, 1/3)$, and set $K_{\perp} = t_q(1 - \delta')\sqrt{1 - \rho_{\perp}^2/q}, \Delta_{\perp} = \max\{\nu - k_{\nu}(\delta'/2), k_{\nu}(1 - \delta'/2) - \nu\}$. Then, if $\rho_{\perp} > K_{\perp}$, $R(b) = M_{\perp} b$ is $(\varepsilon_{\perp}, 3\delta')$-DP for the residual inputs in Definition 4 where*

$$\varepsilon_{\perp} = \frac{\nu + \Delta_{\perp}}{2} \left(\frac{1}{\rho_{\perp} - K_{\perp}} + \log(\rho_{\perp} - K_{\perp}) - 1 \right) - (q - 1 + \Delta_{\perp}/2) \log(\rho_{\perp} - K_{\perp}).$$

Proof. We treat the two cases separately.

Case 1: $\beta_{\perp} = 0$**.** We defined $\|b\| = \beta_{\perp}$ therefore $\beta_{\perp} = 0$ means $b = 0$. Thus, $R(b) = M_{\perp} b M_{\perp} b' = R(b') = 0$.

Case 2: $\beta_\perp > 0$**.** Let $G = U^\top Z$ and recall that $H = \mathrm{rowspan}(G)$ and $Z_\perp = ZP_H^\perp$.

$$\mathcal{L}(M_\perp \mid G, M_\parallel, Y) \;=\; \mathcal{L}(M_\perp \mid G),$$

as $M_\perp \perp M_\parallel | G$.

Therefore it suffices to prove that, conditional on G, the map $b \mapsto M_\perp b$ is $(\varepsilon_\perp, \delta_\perp)$-DP.

Fix G (and hence, fix H and the projector $P_H^\perp$). Using the decomposition $Z = UG + U_\perp W$ and the fact that $GP_H^\perp = 0$, we have

$$Z_\perp \;=\; ZP_H^\perp \;=\; (UG + U_\perp W)P_H^\perp \;=\; U_\perp\, W\, P_H^\perp.$$

Hence

$$M_\perp \;=\; Z_\perp Z_\perp^\top \;=\; U_\perp\,(WP_H^\perp)(WP_H^\perp)^\top\, U_\perp^\top.$$

Consequently, for any $b \in \mathbb{R}^d$,

$$R(b) = M_\perp b \;=\; U_\perp\,(WP_H^\perp)(WP_H^\perp)^\top\, \underbrace{(U_\perp^\top b)}_{=:\, b_\perp}. \tag{22}$$

Step 1: identify the residual mechanism. Conditional on G, the projector $P_H^\perp$ is deterministic and W remains a Gaussian matrix with i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries. Conditional on G, the subspace $H = \mathrm{rowspan}(G)$ is fixed, and hence the projector $P_H^\perp$ is deterministic. Choose an orthonormal basis $Q_\perp \in \mathbb{R}^{r\times(r-p)}$ for $H^\perp$, so that

$$P_H^\perp = Q_\perp Q_\perp^\top.$$

Then

$$WP_H^\perp = WQ_\perp Q_\perp^\top.$$

Since W has i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries and $Q_\perp$ has orthonormal columns, the matrix $WQ_\perp \in \mathbb{R}^{(d-s)\times(r-p)}$ has i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries. Moreover,

$$(WP_H^\perp)(WP_H^\perp)^\top = (WQ_\perp Q_\perp^\top)(Q_\perp Q_\perp^\top W^\top) = (WQ_\perp)(WQ_\perp)^\top.$$

Therefore, conditional on G, the random matrix

$$\widetilde{M}_\perp \;:=\; (WP_H^\perp)(WP_H^\perp)^\top$$

has the same distribution as a (scaled) Wishart matrix in dimension $(d-s)$ with $(r-p)$ degrees of freedom, where $p = \dim(H) = \mathrm{rank}(G)$.

Step 2: apply the vector-DP guarantee and post-processing. Consider the "core" residual mechanism

$$\widetilde{R}(b_\perp) \;:=\; \widetilde{M}_\perp\, b_\perp \qquad \in \mathbb{R}^{d-s}.$$

By Theorem 1, we obtain that conditional on G, the map $b_\perp \mapsto \widetilde{R}(b_\perp)$ is

$$\big(\varepsilon_{\mathrm{vec}}(\rho_\perp; d-s, r-p),\ \delta_{\mathrm{vec}}(\rho_\perp; d-s, r-p)\big)\text{-DP}.$$

Finally, equation 22 shows that $R(b)$ is obtained from $\widetilde{R}(b_\perp)$ by applying the deterministic linear map $x \mapsto U_\perp x$ (given G). Since differential privacy is preserved under post-processing, it follows that conditional on G, the map $b \mapsto R(b) = M_\perp b$ is $(\varepsilon_\perp, \delta_\perp)$-DP. The exact $(\varepsilon_\perp, \delta_\perp)$ is then obtained by instantiating Theorem 1 with a Wishart matrix of dimension $(d-s)\times(r-p)$ and alignment parameter $\rho_\perp$. □

Lemma 15. *Fix a unit vector $u \in \mathbb{R}^d$ independent of Z, and write $u_\parallel = P_S u, u_\perp = P_{S^\perp} u$. Define the directional effective rank*

$$r_{\mathrm{eff}}(u) \coloneqq r\left\|u_\parallel\right\|^2 + p\left\|u_\perp\right\|^2.$$

Fix $\delta_s \in (0,1)$. Conditional on H, with probability at least $1-\delta_s$ over the draw of Z,

$$\left\|M_\parallel u\right\| \le \sqrt{2}\,\sigma_M^2\left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln(3/\delta_s)}\right)\left(\sqrt{r_{\mathrm{eff}}(u)} + \sqrt{2\ln(3/\delta_s)}\right) \coloneqq \Gamma_{\delta_s}(u). \tag{23}$$

Proof. Recall that $S = \mathrm{span}(V_{-j})$, let U be an orthonormal basis matrix for S, and write $G = U^\top Z, H = \mathrm{rowspan}(G)$, and $p = \dim(H)$. Let P_H be the orthogonal projector onto H. Since $u = u_\parallel + u_\perp$ and $Z^\top u_\parallel \in H$, we have

$$\begin{aligned}\left\|M_\parallel u\right\| &= \left\|ZP_H Z^\top \left(u_\parallel + u_\perp\right)\right\| \\ &\le \|Z\|_{\mathrm{op}} \left(\left\|Z^\top u_\parallel\right\| + \left\|P_H Z^\top u_\perp\right\|\right).\end{aligned}$$

We now bound the three terms in equation E.1. First, a standard Gaussian operator-norm bound gives, for every $t \ge 0$,

$$\Pr\left(\|Z\|_{\mathrm{op}} > \sigma_M\left(\sqrt{d} + \sqrt{r} + t\right)\middle| H\right) \le e^{-t^2/2}.$$

Likewise,

$$\Pr\left(\left\|Z^\top u_\parallel\right\| > \sigma_M \left\|u_\parallel\right\| \left(\sqrt{r} + t\right)\middle| H\right) \le e^{-t^2/2}.$$

We note that although Z is not conditionally i.i.d. Gaussian given H, the norms $\|Z\|_{\mathrm{op}}$ and $\|Z^\top u_\parallel\|$ retain their usual Gaussian tail bounds as before conditioning on H.

Because $W = U_\perp^\top Z$ is independent of $G = U^\top Z$, it is also independent of $H = \mathrm{rowspan}(G)$. Moreover, $Z^\top u_\perp = W^\top U_\perp^\top u_\perp$. Therefore, conditional on H,

$$\frac{\left\|P_H Z^\top u_\perp\right\|}{\sigma_M \left\|u_\perp\right\|} \sim \chi_p.$$

The standard χ_p tail bound then gives, for every $t \ge 0$,

$$\Pr\left(\left\|P_H Z^\top u_\perp\right\| > \sigma_M \left\|u_\perp\right\| \left(\sqrt{p} + t\right)\middle| H\right) \le e^{-t^2/2}.$$

Set $t = \sqrt{2\ln(3/\delta_s)}$, combine this bound with the two preceding Gaussian bounds, and applying a union bound to obtain, with conditional probability at least $1 - \delta_s$,

$$\left\|M_\parallel u\right\| \le \sigma_M^2 \left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln(3/\delta_s)}\right)\left[\left(\sqrt{r} + \sqrt{2\ln(3/\delta_s)}\right)\left\|u_\parallel\right\| + \left(\sqrt{p} + \sqrt{2\ln(3/\delta_s)}\right)\left\|u_\perp\right\|\right].$$

Finally, since $\|u\| = 1$, $u_\parallel \perp u_\perp$, and substituting r_{eff} we observe

$$\left(\sqrt{r} + \sqrt{2\ln(3/\delta_s)}\right)\left\|u_\parallel\right\| + \left(\sqrt{p} + \sqrt{2\ln(3/\delta_s)}\right)\left\|u_\perp\right\| \le \sqrt{2}\left(\sqrt{r_{\mathrm{eff}}(u)} + \sqrt{2\ln(3/\delta_s)}\right).$$

Hence

$$\left\|M_\parallel u\right\| \le \sqrt{2}\,\sigma_M^2 \left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln(3/\delta_s)}\right)\left(\sqrt{r_{\mathrm{eff}}(u)} + \sqrt{2\ln(3/\delta_s)}\right)$$

□

Lemma 16. *Fix $\delta_s \in (0,1)$ and $\delta_\parallel \in (0,1)$. Consider the mechanism $C(v) = M_\parallel v + \xi$, where $\xi \sim \mathcal{N}(0, \sigma_G^2 I_d)$ is independent of $M_\parallel$. Suppose neighbouring inputs satisfy $\|v - v'\| \le \Delta_v$. Then, C is $(\varepsilon_\parallel, \delta_\parallel + \delta_s)$-DP, where*

$$\varepsilon_\parallel = \frac{2\sigma_M^2\left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln\frac{3}{\delta_s}}\right)\left(\sqrt{2r(1-\beta_\perp^2)(1-\rho_\parallel) + 2p\beta_\perp^2(1-\rho_\perp)} + \sqrt{2\ln\frac{3}{\delta_s}}\Delta_v\right)}{\sigma_G}\sqrt{2\ln\left(\frac{1}{\delta_\parallel}\right)}.$$

Proof. Fix any neighbouring $v \sim v'$ and define $u = \frac{v - v'}{\|v - v'\|}$. Let $\mathcal{E}_{u,\delta_s}$ be the event

$$\mathcal{E}_{u,\delta_s} := \left\{\left\|M_\parallel u\right\| \le \Gamma_{\delta_s}\right\}.$$

By Lemma 15, we have $\Pr\left(\mathcal{E}_{u,\delta_s}\right) \ge 1 - \delta_s$.

Conditional on $M_\parallel$, the outputs are Gaussians

$$C(v) \mid M_\parallel \sim \mathcal{N}\left(M_\parallel v, \sigma_G^2 I_d\right), \qquad C(v') \mid M_\parallel \sim \mathcal{N}\left(M_\parallel v', \sigma_G^2 I_d\right),$$

whose means differ by

$$M_\parallel (v - v') = \|v - v'\| M_\parallel u.$$

On $\mathcal{E}_{u,\delta_s}$, we have

$$\left\|M_\parallel \|v - v'\| u\right\| \le \Gamma_{\delta_s} \|v - v'\| = \sqrt{2}\sigma_M^2 \left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln\frac{3}{\delta_s}}\right)\left(\sqrt{r_{\text{eff}}(u)\|v - v'\|^2} + \sqrt{2\ln\frac{3}{\delta_s}}\|v - v'\|\right) \tag{24}$$

Let $h = v - v'$. We can then decompose h into the space S and $S^\perp$. Specifically, let

$$h = (a - a') + (b - b') \tag{25}$$

where $a, a' \in S$ and $b, b' \in S^\perp$.

Then,

$$\begin{aligned} r_{\text{eff}}(u) &= r\|P_S u\|^2 + p\|P_{S^\perp} u\|^2 \\ &= r\left\|\frac{\|a - a'\|^2}{\|h\|^2}\right\| + p\left\|\frac{\|b - b'\|^2}{\|h\|^2}\right\| \\ &= \frac{r\|a - a'\|^2 + p\|b - b'\|^2}{\|h\|^2} \end{aligned}$$

Hence, for all v v' such that the corresponding neighboring datasets V, V' satisfying Definition 4 with parameters $\beta_\perp, \rho_\perp, \rho_\parallel$, the following holds,

$$r_{\text{eff}}(u)\|v - v'\|^2 = r\|a - a'\|^2 + p\|b - b'\|^2 \le 2r(1 - \beta_\perp^2)(1 - \rho_\parallel) + 2p\beta_\perp^2(1 - \rho_\perp) \tag{26}$$

Substituting Equation (26) into Equation (24), we get,

$$\left\|M_\parallel(v - v')\right\| \le \sqrt{2}\sigma_M^2 \left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln\frac{3}{\delta_s}}\right)\left(\sqrt{2r(1 - \beta_\perp^2)(1 - \rho_\parallel) + 2p\beta_\perp^2(1 - \rho_\perp)} + \sqrt{2\ln\frac{3}{\delta_s}}\Delta_v\right) \tag{27}$$

Therefore, on $\mathcal{E}_{u,\beta}$ the standard Gaussian mechanism analysis implies that $C(\cdot)$ is $(\varepsilon_\parallel, \delta_\parallel)$-DP with

$$\varepsilon_\parallel = \frac{2\sigma_M^2\left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln\frac{3}{\delta_s}}\right)\left(\sqrt{2r(1 - \beta_\perp^2)(1 - \rho_\parallel) + 2p\beta_\perp^2(1 - \rho_\perp)} + \sqrt{2\ln\frac{3}{\delta_s}}\Delta_v\right)}{\sigma_G}\sqrt{2\ln\left(\frac{1}{\delta_\parallel}\right)}.$$

Finally, remove the conditioning: for any measurable set $S \subseteq \mathbb{R}^d$,

$$\begin{aligned} \Pr(C(v) \in S) &\le \Pr(C(v) \in S \mid \mathcal{E}_{u,\beta}) \Pr(\mathcal{E}_{u,\beta}) + \Pr\left(\mathcal{E}_{u,\beta}^c\right) \\ &\le \left(e^{\varepsilon_\parallel} \Pr(C(v') \in S \mid \mathcal{E}_{u,\beta}) + \delta_\parallel\right) \Pr(\mathcal{E}_{u,\beta}) + \Pr\left(\mathcal{E}_{u,\beta}^c\right) \\ &\le e^{\varepsilon_\parallel} \Pr(C(v') \in S) + \delta_\parallel + \Pr\left(\mathcal{E}_{u,\beta}^c\right) \\ &\le e^{\varepsilon_\parallel} \Pr(C(v') \in S) + \delta_\parallel + \delta_s. \end{aligned}$$

□

Theorem 4. *Fix $j \in [n]$. Let $V \sim^{(j)} V'$ with parameters $(\rho_\parallel, \beta_\perp, \rho_\perp)$ defined in Definition 4. Fix $\delta_\parallel, \delta_s \in (0, 1)$, and choose $\delta' > 0$ as in Lemma 2 to obtain $(\varepsilon_\perp, \delta_\perp)$. Let $p = \operatorname{rank}(U^\top Z)$ and define Γ_{δ_s} as above, and $\varepsilon_\parallel$ as in Equation* (4)*. Then the mechanism $\mathcal{A}_1(V) = MV + \Xi$ is $(\varepsilon_{\text{one}}, \delta_{\text{one}})$-DP for inputs $V \sim_A^{(j)} V'$, with $\varepsilon_{\text{one}} \le \varepsilon_\parallel + \varepsilon_\perp$ and $\delta_{\text{one}} \le \delta_\parallel + \delta_\perp + \delta_s$.*

Proof. We can write the the mechanism $\mathcal{A}(V) = MV + \Xi$ as

$$\begin{aligned}(X, Y) &= (Mv_j + \xi_j \in \mathbb{R}^d, [Mv_k + \xi_k]_{k\neq j} \in \mathbb{R}^{d\times(n-1)}) \\ &= (M_\parallel v_j + \xi_j + M_\perp b, \left[M_\parallel v_k + \xi_k\right]_{k\neq j} \in \mathbb{R}^{d\times(n-1)}).\end{aligned}$$

So if we define

$$\begin{aligned}R(b) &= M_\perp b \\ C(v) &= M_\parallel v_j\end{aligned}$$

we know by Lemma 14, conditional on H, the residual randomness $M_\perp$ is independent of Y. Therefore, Lemma 2 tells us that conditional on (H, Y) R is $(\varepsilon_\perp, \delta_\perp)$-DP.

For the correlated term C, Lemma 16 establishes $(\varepsilon_\parallel, \delta_\parallel)$-differential privacy using the Gaussian mechanism with the directional sensitivity bound from Lemma 15, conditional on the good event of $M_\parallel$.

By composition and the postprocessing Lemma this tells us that conditioned on H and $\|M_\parallel u\| \leq \Gamma_\beta$ (X, Y) is $(\varepsilon_\parallel + \varepsilon_\perp, \delta_\parallel + \delta_\perp)$-DP. We can remove the conditioning on $\|M_\parallel u\| \leq \Gamma_\beta$ by adding an additional δ_s to our final δ. The conditioning on H can be removed because H is a measurable function of (Z, V_{-j}). Under $V \sim_A^{(j)} V'$, we have $V_{-j} = V'_{-j}$, so the law of H is the same under V and V'. Therefore conditioning on H does not affect the DP comparison. □

Notation for Theorem 1. For $\rho \in (0, 1]$, integers $d' \geq 1$, $r' \geq 1$ and $\eta \in (0, 1/3)$, write $\nu := d' + r' - 1$ and

$$K(\rho; r', \eta) := \sqrt{\tfrac{1-\rho^2}{r'}}\, t_{r'}(1-\eta), \qquad \Delta(d', r', \eta) := \max\{\nu - \kappa_\nu(\tfrac{\eta}{2}),\ \kappa_\nu(1-\tfrac{\eta}{2}) - \nu\},$$

where t_ℓ and κ_ℓ are the quantile functions of the Student-t and χ^2 distributions with ℓ degrees of freedom. Whenever $\rho > K := K(\rho; r', \eta)$, set $\Delta := \Delta(d', r', \eta)$ and

$$\varepsilon_{\mathrm{vec}}(\rho; d', r', \eta) := \frac{\nu + \Delta}{2}\Big(\frac{1}{\rho - K} + \log(\rho - K) - 1\Big) - \Big(r' - 1 + \frac{\Delta}{2}\Big)\log(\rho - K).$$

In this notation Theorem 1 reads: let $\widetilde{Z} \in \mathbb{R}^{d'\times r'}$ have i.i.d. $\mathcal{N}(0, \sigma^2)$ entries and $\widetilde{M} := \widetilde{Z}\widetilde{Z}^\top$; if $\rho > K(\rho; r', \eta)$, then for all unit vectors $v, v' \in \mathbb{R}^{d'}$ with $\langle v, v'\rangle \geq \rho$ and all measurable $E' \subseteq \mathbb{R}^{d'}$,

$$\Pr(\widetilde{M}v \in E') \ \leq\ e^{\varepsilon_{\mathrm{vec}}(\rho; d', r', \eta)}\ \Pr(\widetilde{M}v' \in E') + 3\eta.$$

E.2 Proofs in the Small r Regime

For

$$\begin{aligned}Y &= M(V + \sigma E) \\ Y' &= M(V' + \sigma E) \\ M &= \sum_i^r z_i z_i^\top \text{ with } z_i \sim \mathcal{N}(0, I_d) \\ \Delta V &= V - V'\end{aligned}$$

let P be the law of Y and Q the law of Y'. Then because P and Q are mutually absolute continuous we are able to define the the density ratio.

Lemma 17. *Let $P(\cdot \mid M)$ and $Q(\cdot \mid M)$ denote the conditional laws of*

$$Y = MV + \sigma ME, \qquad Y' = MV' + \sigma ME,$$

respectively, where the entries of E are i.i.d. standard Gaussians. Define

$$L_M(y) := \frac{dP(\cdot \mid M)}{dQ(\cdot \mid M)}(y), \qquad \Delta V := V - V', \qquad \rho(M) := \frac{\|P_M \Delta V\|_F^2}{\sigma^2},$$

where $P_M := M(M^2)^\dagger M$ is the orthogonal projector onto $\mathrm{range}(M)$*. Then, conditional on M,*

$$\log L_M(Y) \sim \mathcal{N}\left(\frac{\rho(M)}{2}, \rho(M)\right), \qquad Y \sim P(\cdot \mid M),$$

whereas

$$\log L_M(Y') \sim \mathcal{N}\left(-\frac{\rho(M)}{2}, \rho(M)\right), \qquad Y' \sim Q(\cdot \mid M).$$

Consequently,

$$\Pr_{Y \sim P(\cdot|M)}(|\log L_M(Y)| > \varepsilon \mid M) = T(\varepsilon; \rho(M)),$$

and the same identity holds with $Y' \sim Q(\cdot \mid M)$.

Proof. Since $Y|M$ is a Gaussian and any affine function of a Gaussian is Gaussian (and the log likelihood ratio is affine) we have that $\log L_M(Y)|M$ is Gaussian. So we only need to determine its mean and variance. For $Y|M \sim \mathcal{N}(\mu, \Sigma)$ and $Y'|M \sim \mathcal{N}(\mu', \Sigma)$ we have that

$$\log L_M(y) = (\mu - \mu')^\top \Sigma^\dagger (y - \frac{\mu + \mu'}{2})$$

where $\Sigma^\dagger$ is the pseudoinverse.

Mean:

$$\begin{aligned}
\mathbb{E}[\log L_M(Y)|M] &= \mathbb{E}[(\mu - \mu')^\top \Sigma^\dagger (Y - \frac{\mu + \mu'}{2})|M] \\
&= (\mu - \mu')^\top \Sigma^\dagger (\mathbb{E}[Y|M] - \frac{\mu + \mu'}{2}) \\
&= \frac{1}{2}(\mu - \mu')^\top \Sigma^\dagger (\mu - \mu')
\end{aligned}$$

where the last step follows as by the definition of the log-likelihood that Y is distributed by the nominator. (The inverse log likelihood would lead to a minus sign in the mean)

Variance:

$$\begin{aligned}
\mathrm{Var}(\log L_M(Y)|M) &= \mathrm{Var}((\mu - \mu')^\top \Sigma^\dagger Y|M) = \mathbb{E}[((\mu - \mu')^\top \Sigma^\dagger Y)^2|M] - \mathbb{E}[(\mu - \mu')^\top \Sigma^\dagger Y|M]^2 \\
&= \mathbb{E}[(\mu - \mu')^\top \Sigma^\dagger Y Y^\top \Sigma^\dagger (\mu - \mu')|M] - ((\mu - \mu')^\top \Sigma^\dagger \mu)((\mu - \mu')^\top \Sigma^\dagger \mu)^\top \\
&= (\mu - \mu')^\top \Sigma^\dagger \mathbb{E}[YY^\top|M] \Sigma^\dagger (\mu - \mu') - (\mu - \mu')^\top \Sigma^\dagger \mu\mu^\top \Sigma^\dagger (\mu - \mu') \\
&= (\mu - \mu')^\top \Sigma^\dagger (\mathbb{E}[YY^\top|M] - \mu\mu^\top) \Sigma^\dagger (\mu - \mu') \\
&= (\mu - \mu')^\top \Sigma^\dagger \mathrm{Var}(Y|M) \Sigma^\dagger (\mu - \mu') \\
&= (\mu - \mu')^\top \Sigma^\dagger \Sigma \Sigma^\dagger (\mu - \mu') = (\mu - \mu')^\top \Sigma^\dagger (\mu - \mu')
\end{aligned}$$

Now let's recall that

$$\begin{aligned}
Y &= MV + \sigma ME \\
Y' &= MV' + \sigma ME
\end{aligned}$$

so the jth column is distributed as

$$\begin{aligned}
Y_{:,j}|M &\sim \mathcal{N}(MV_{:,j}, \sigma^2 M^2) \\
Y'_{:,j}|M &\sim \mathcal{N}(MV'_{:,j}, \sigma^2 M^2)
\end{aligned}$$

and the columns are independent given M. So if we define $y = \mathrm{vec}(Y)$ (stack columns into one vector) we have that $\mu = \mathrm{vec}(MV), \mu' = \mathrm{vec}(MV')$ and because columns are independent given M, $\mathrm{Cov}(Y_{:,i}, Y_{:,j}|M) = 0$ for $i \neq j$ which means

$$\Sigma \;=\; \mathrm{Cov}(y \mid M) \;=\; \begin{pmatrix} \sigma^2 M^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 M^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 M^2 \end{pmatrix} \;=\; \sigma^2 \,\mathrm{diag}\big(M^2, \ldots, M^2\big) = \sigma^2 \big(I_k \otimes M^2\big).$$

Recall we need the pseudoinverse of our variance, for which we will use two identities:

$$(\alpha A)^\dagger = \frac{1}{\alpha} A^\dagger$$
$$(A \otimes B)^\dagger = A^\dagger \otimes B^\dagger$$

so we get

$$\Sigma^\dagger = \frac{1}{\sigma^2}(I_k \otimes (M^2)^\dagger).$$

Then by the Kronecker-Vec Identity we have

$$(\Delta\mu)^\top \Sigma^\dagger (\Delta\mu) \;=\; \frac{1}{\sigma^2}\,\mathrm{tr}\big((M\Delta V)^\top (M^2)^\dagger (M\Delta V)\big)\,.$$

Using $P_M = M(M^2)^\dagger M$ and $P_M^2 = P_M$, we get

$$(\Delta\mu)^\top \Sigma^\dagger (\Delta\mu) = \frac{1}{\sigma^2}\,\mathrm{tr}\big(\Delta V^\top P_M \Delta V\big) = \frac{\|P_M \Delta V\|_F^2}{\sigma^2}.$$

Plugging this into the mean and variance formulas and using

$$\rho(M) = \frac{\|P_M \Delta V\|_F^2}{\sigma^2},$$

gives the two conditional privacy-loss distributions

$$\log L_M(Y) \mid M \sim \mathcal{N}\left(\frac{\rho(M)}{2}, \rho(M)\right), \qquad Y \sim P(\cdot \mid M),$$

and

$$\log L_M(Y') \mid M \sim \mathcal{N}\left(-\frac{\rho(M)}{2}, \rho(M)\right), \qquad Y' \sim Q(\cdot \mid M).$$

Consequently, under $Y \sim P(\cdot \mid M)$,

$$\begin{aligned}
\Pr\left(|\log L_M(Y)| > \varepsilon \mid M\right) &= \Phi\left(\frac{-\varepsilon - \rho(M)/2}{\sqrt{\rho(M)}}\right) + 1 - \Phi\left(\frac{\varepsilon - \rho(M)/2}{\sqrt{\rho(M)}}\right) \\
&= \Phi\left(\frac{-\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) + 1 - \Phi\left(\frac{\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) \\
&= T(\varepsilon; \rho(M)),
\end{aligned}$$

where the second equality uses $\Phi(-x) = 1 - \Phi(x)$. Similarly, under $Y' \sim Q(\cdot \mid M)$,

$$\begin{aligned}
\Pr\left(|\log L_M(Y')| > \varepsilon \mid M\right) &= \Phi\left(\frac{-\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) + 1 - \Phi\left(\frac{\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) \\
&= T(\varepsilon; \rho(M)).
\end{aligned}$$

□

Lemma 18 (Tail bound for Haar projection)**.** *Let $r \in \{1, \ldots, d-1\}$ and let $Z \in \mathbb{R}^{d\times r}$ have i.i.d. $\mathcal{N}(0,1)$ entries. Let P_M denote the orthogonal projector onto $\mathrm{col}(Z)$ (equivalently onto $\mathrm{col}(M)$ for $M = ZZ^\top$). Fix a deterministic matrix $\Delta V \in \mathbb{R}^{d\times n}$ of rank $s \geq 1$, and write $\|\cdot\|_F$ for the Frobenius norm. Then for every $\alpha \in (0,1)$,*

$$\Pr\left(\frac{\|P_M \Delta V\|_F^2}{\|\Delta V\|_F^2} > \alpha\right) \;\leq\; s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right],$$

where $I_\alpha(a,b)$ is the regularized incomplete beta function (the $\mathrm{Beta}(a,b)$ CDF).

Remark 1. *In our setting the matrix Z is generated with i.i.d. entries $Z_{ij} \sim \mathcal{N}(0, 1/r)$. This differs from the standard $Z_{ij} \sim \mathcal{N}(0,1)$ only by a scalar factor: $Z = \frac{1}{\sqrt{r}} G$ with $G_{ij} \sim \mathcal{N}(0,1)$. Since scaling by a nonzero constant does not change the column space, $\mathrm{col}(Z) = \mathrm{col}(G)$, and hence the orthogonal projector P_M onto $\mathrm{col}(Z)$ has the same distribution. Therefore Lemma 18 applies unchanged.*

Proof. Because Z has i.i.d. standard normal entries, its law is orthogonally invariant:

$$UZ \overset{d}{=} Z \qquad \text{for all } U \in O(d).$$

Therefore $\mathrm{col}(Z)$ has a rotation-invariant distribution on the Grassmannian $\mathrm{Gr}(d, r)$ (see Definitions 8 and 11), hence it is uniform. Moreover, in the (thin) QR decomposition $Z = QR$, the factor $Q \in V_{d,r}$ inherits the same invariance

$$UQ \overset{d}{=} Q \qquad \text{for all } U \in O(d),$$

and is thus Haar-uniform on the Stiefel manifold in the sense of Definition 10. Since $P_M = QQ^\top$ and Q is Haar-uniform on $V_{d,r}$, for any fixed $U \in O(d)$ we have

$$UP_MU^\top \;=\; UQQ^\top U^\top \;=\; (UQ)(UQ)^\top.$$

By Haar-uniformity, $UQ \overset{d}{=} Q$, hence

$$UP_MU^\top \;\overset{d}{=}\; QQ^\top \;=\; P_M.$$

Thus the law of P_M is invariant under conjugation by orthogonal matrices. This is exactly the notion of Haar-uniformity for rank-r orthogonal projectors.

Assume ΔV has rank 1, so $\Delta V = uw^\top$ with $\|u\|_2 = 1$. Then

$$\frac{\|P_M \Delta V\|_F^2}{\|\Delta V\|_F^2} = \frac{\|P_M uw^\top\|_F^2}{\|uw^\top\|_F^2} = \frac{\|P_M u\|_2^2 \|w\|_2^2}{\|u\|_2^2 \|w\|_2^2} = \|P_M u\|_2^2.$$

Let

$$P_0 \;\coloneqq\; \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Since every rank-r orthogonal projector is an orthogonal conjugate of P_0, and P_M is Haar-uniform, we may write $P_M \overset{d}{=} UP_0U^\top$ with $U \sim \mathrm{Haar}(O(d))$. Then $y \coloneqq U^\top u$ is uniform on the unit sphere S^{d-1}, and

$$X := \|P_M u\|_2^2 = \sum_{i=1}^{r} y_i^2.$$

Let $g \sim \mathcal{N}(0, I_d)$ and note $g/\|g\|_2$ is uniform on S^{d-1} . Write $g = (g_{1:r}, g_{r+1:d})$. Then

$$X \overset{d}{=} \frac{\sum_{i=1}^{r} g_i^2}{\sum_{i=1}^{d} g_i^2} = \frac{U}{U+V}, \qquad U \sim \chi_r^2,\ V \sim \chi_{d-r}^2 \text{ independent.}$$

Hence $X \sim \mathrm{Beta}\big(\frac{r}{2}, \frac{d-r}{2}\big)$. Therefore,

$$\Pr(X > \alpha) = 1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right).$$

For a ΔV with general rank, by SVD we obtain $\Delta V = \sum_{j=1}^{s} \sigma_j u_j v_j^\top$ with orthonormal $\{u_j\}_{j=1}^s$. Using $P_M^\top P_M = P_M$ and orthonormality,

$$\|P_M \Delta V\|_F^2 = \sum_{j=1}^{s} \sigma_j^2 \|P_M u_j\|_2^2, \qquad \|\Delta V\|_F^2 = \sum_{j=1}^{s} \sigma_j^2.$$

Define weights $w_j := \sigma_j^2 / \sum_{\ell=1}^{s} \sigma_\ell^2$ and $X_j := \|P_M u_j\|_2^2 \in [0, 1]$. Then

$$\frac{\|P_M \Delta V\|_F^2}{\|\Delta V\|_F^2} = \sum_{j=1}^{s} w_j X_j.$$

If $\sum_{j=1}^{s} w_j X_j > \alpha$ and $\sum_j w_j = 1$ with $w_j \geq 0$, then necessarily $\max_{1 \leq j \leq s} X_j > \alpha$ (otherwise all $X_j \leq \alpha$ would imply the weighted average is $\leq \alpha$). Thus,

$$\Pr\left(\sum_{j=1}^{s} w_j X_j > \alpha\right) \leq \Pr\left(\max_{1 \leq j \leq s} X_j > \alpha\right) \leq \sum_{j=1}^{s} \Pr(X_j > \alpha).$$

Finally, each X_j has the same marginal law as in Step 2 because u_j is a fixed unit vector and P_M is Haar, so $\Pr(X_j > \alpha) = 1 - I_\alpha\big(\frac{r}{2}, \frac{d-r}{2}\big)$. Therefore,

$$\Pr\left(\frac{\|P_M \Delta V\|_F^2}{\|\Delta V\|_F^2} > \alpha\right) \le s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right],$$

as claimed. □

Theorem 5 (Privacy of M2 in the small r-regime)**.** *Fix $V, V' \in \mathbb{R}^{d\times n}$ and write $s = \min(n, d)$. Then for any $\alpha \in (0,1)$ and any $\varepsilon > 0$, $0 < r < d$, the mechanism $\mathcal{A}_2$ is $(\varepsilon, \delta_\alpha(\varepsilon))$-DP with*

$$\delta_\alpha(\varepsilon) \le T\left(\varepsilon; \frac{\alpha\,\|\Delta V\|_F^2}{\sigma_G^2}\right) + s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right],$$

where $I_\alpha(a,b)$ is the regularized incomplete Beta function. Moreover, the same privacy guarantee holds for $M'(V + \Xi)$, where M' is obtained by orthonormalizing M.

Proof. First note that

$$\begin{aligned}\mathbb{P}(Y \in A) &= \mathbb{E}_M[\mathbb{P}(Y \in A|M)] = \mathbb{E}_M[\mathbb{P}(Y \in A|M)\mathbf{1}_{\{M\in\mathcal{G}_\alpha\}}] + \mathbb{E}_M[\mathbb{P}(Y \in A|M)\mathbf{1}_{\{M\notin\mathcal{G}_\alpha\}}] \\ &\le \mathbb{E}_M[\mathbb{P}(Y \in A|M)\mathbf{1}_{\{M\in\mathcal{G}_\alpha\}}] + \delta_M\end{aligned}$$

where $\delta_M \coloneqq \mathbb{P}(M \notin \mathcal{G}_\alpha)$ and $\mathcal{G}_\alpha$ will be defined later. This means we can analyse $\mathbb{E}_M[\mathbb{P}(Y \in A|M)\mathbf{1}_{\{M\in\mathcal{G}_\alpha\}}]$ separately and find ε, δ so that

$$\mathbb{E}\big[\mathbb{P}(Y \in A \mid M)\,\mathbf{1}_{\{M\in G_\alpha\}}\big] \;\le\; e^\varepsilon\,\mathbb{E}\big[\mathbb{P}(Y' \in A \mid M)\,\mathbf{1}_{\{M\in G_\alpha\}}\big] \;+\; \delta_E(\varepsilon,\alpha)\,\mathbb{P}(M \in G_\alpha).$$

Bound $\mathbb{P}(M \in G_\alpha) \le 1$ and note that

$$\mathbb{E}\big[\mathbb{P}(Y' \in A \mid M)\,\mathbf{1}_{\{M\in G_\alpha\}}\big] \;\le\; \mathbb{E}[\mathbb{P}(Y' \in A \mid M)] \;=\; \mathbb{P}(Y' \in A).$$

Therefore,

$$\mathbb{E}\big[\mathbb{P}(Y \in A \mid M)\,\mathbf{1}_{\{M\in G_\alpha\}}\big] \;\le\; e^\varepsilon\,\mathbb{P}(Y' \in A) \;+\; \delta_E(\varepsilon,\alpha).$$

Combine with the δ_M bound for the complement to obtain

$$\mathbb{P}(Y \in A) \;\le\; e^\varepsilon\,\mathbb{P}(Y' \in A) \;+\; \delta_E(\varepsilon,\alpha) \;+\; \delta_M.$$

Lemma 18 gives us a bound on δ_M. For fixed M, let P_M and Q_M denote the conditional laws of the mechanism under V and V', respectively, and write

$$L_M(y) := \frac{dP_M}{dQ_M}(y), \qquad \rho(M) := \frac{\|P_M \Delta V\|_F^2}{\sigma_G^2}.$$

For fixed M, define the conditional good-output set

$$\mathcal{Y}_\varepsilon(M) := \{y : |\log L_M(y)| \le \varepsilon\}.$$

On $\mathcal{Y}_\varepsilon(M)$, we have the pointwise bound $L_M(y) \le e^\varepsilon$.

By Lemma 17, if $Y \sim P_M$, then

$$\log L_M(Y) \mid M \sim \mathcal{N}\left(\frac{\rho(M)}{2}, \rho(M)\right),$$

whereas if $Y' \sim Q_M$, then

$$\log L_M(Y') \mid M \sim \mathcal{N}\left(-\frac{\rho(M)}{2}, \rho(M)\right).$$

Although these two privacy-loss distributions have opposite means, their two-sided tails are equal. Indeed, using $\Phi(-x) = 1 - \Phi(x)$,

$$\begin{aligned}\Pr_{Y\sim P_M}(Y \notin \mathcal{Y}_\varepsilon(M) \mid M) &= \Phi\left(\frac{-\varepsilon - \rho(M)/2}{\sqrt{\rho(M)}}\right) + 1 - \Phi\left(\frac{\varepsilon - \rho(M)/2}{\sqrt{\rho(M)}}\right) \\ &= \Phi\left(\frac{-\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) + 1 - \Phi\left(\frac{\varepsilon + \rho(M)/2}{\sqrt{\rho(M)}}\right) \\ &= T(\varepsilon; \rho(M)).\end{aligned}$$

The same two-sided tail value is obtained under Q_M.

Now fix a parameter $\alpha \in (0, 1]$ and define the alignment-good event

$$G_\alpha \;:=\; \left\{\, M : \|P_M \Delta V\|_F^2 \le \alpha \,\|\Delta V\|_F^2 \,\right\}.$$

On G_α,

$$\rho(M) \le \bar{\rho} := \frac{\alpha \|\Delta V\|_F^2}{\sigma_G^2}.$$

Since $T(\varepsilon; \rho)$ is nondecreasing in ρ, for every $M \in G_\alpha$,

$$\Pr_{Y\sim P_M}(Y \notin \mathcal{Y}_\varepsilon(M) \mid M) \le T(\varepsilon; \bar{\rho}).$$

Therefore, for every measurable set A and every $M \in G_\alpha$,

$$\begin{aligned}P_M(A) &= P_M(A \cap \mathcal{Y}_\varepsilon(M)) + P_M(A \cap \mathcal{Y}_\varepsilon(M)^c) \\ &\le e^\varepsilon Q_M(A) + T(\varepsilon; \bar{\rho}).\end{aligned}$$

Averaging this bound over $M \in G_\alpha$, we obtain

$$\begin{aligned}\Pr(Y \in A) &= \mathbb{E}_M[P_M(A)\mathbf{1}_{G_\alpha}] + \mathbb{E}_M[P_M(A)\mathbf{1}_{G_\alpha^c}] \\ &\le e^\varepsilon \mathbb{E}_M[Q_M(A)\mathbf{1}_{G_\alpha}] + T(\varepsilon; \bar{\rho}) + \Pr(M \notin G_\alpha) \\ &\le e^\varepsilon \Pr(Y' \in A) + T(\varepsilon; \bar{\rho}) + \delta_M.\end{aligned}$$

Finally, Lemma 18 gives

$$\delta_M \le s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right],$$

and hence

$$\delta_\alpha(\varepsilon) \le T\left(\varepsilon; \frac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}\right) + s\left[1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right].$$

□

Remark 2 (Orthonormalized projection)**.** *The same proof applies if M is replaced by its orthonormalized version*

$$M' := P_M = M(M^2)^\dagger M,$$

the orthogonal projector onto $\mathrm{col}(M)$*. Indeed, Lemma 17 shows that, conditional on M, the privacy-loss distribution of $M(V + \Xi)$ depends on M only through*

$$\rho(M) = \frac{\|P_M \Delta V\|_F^2}{\sigma_G^2}.$$

If M is replaced by $M' = P_M$, then $(M')^2 = M'$ and $(M')^\dagger = M'$, so the same calculation gives

$$\rho(M') = \frac{\|M'\Delta V\|_F^2}{\sigma_G^2} = \frac{\|P_M \Delta V\|_F^2}{\sigma_G^2} = \rho(M).$$

Moreover, orthonormalization preserves the column space, $\mathrm{col}(M') = \mathrm{col}(M)$*, so the good-projection event and its Haar-projection tail bound Lemma 18 are unchanged. Hence, all following steps in the proof of Theorem 5 holds for M' as well.*

Corollary 1. *Fix $\varepsilon > 0$. Suppose that r satisfies the scaling regime*

$$\log s \;\lesssim\; r \;\ll\; d.$$

Then there exists a choice of α on the order of $\alpha \;\approx\; \frac{r}{d}$, such that the privacy bound from Theorem 5 is strictly smaller than the Gaussian baseline, i.e.,

$$\delta_{\mathrm{ours}}(\varepsilon) \;<\; \delta_{\mathrm{Gauss}}(\varepsilon).$$

Remark 3. *The lower bound $r \gtrsim \log s$ controls the prefactor s in the Beta-tail term (introduced via a union bound over the s sensitive directions). The condition is likely conservative and a sharper control of the s-dimensional subspace could reduce the $\log s$ requirement.*

Proof. Fix $\varepsilon > 0$ and let $\mu = \|\Delta V\|_F^2/\sigma^2$, so that

$$\delta_{\mathrm{Gauss}}(\varepsilon) = T(\varepsilon;\mu) := \Phi\!\left(\frac{-\varepsilon - \mu/2}{\sqrt{\mu}}\right) \;+\; 1 - \Phi\!\left(\frac{\varepsilon - \mu/2}{\sqrt{\mu}}\right).$$

We compare this baseline to the small-r bound from Theorem 5, namely, for any $\alpha \in (0,1)$,

$$\delta_{\mathrm{ours}}(\varepsilon;\alpha) \;\le\; T(\varepsilon;\alpha\mu) + s\left[1 - I_\alpha\!\left(\frac{r}{2}, \frac{d-r}{2}\right)\right].$$

The map $x \mapsto T(\varepsilon;x)$ is continuous and strictly increasing on $x \ge 0$, and satisfies $T(\varepsilon;0) = 0$ and $T(\varepsilon;\mu) = \delta_{\mathrm{Gauss}}(\varepsilon)$. Hence, by the intermediate value theorem, there exists $\alpha_0 \in (0,1)$ such that

$$T(\varepsilon;\alpha_0\mu) \;=\; \frac{1}{2}\,T(\varepsilon;\mu) \;=\; \frac{1}{2}\,\delta_{\mathrm{Gauss}}(\varepsilon). \tag{28}$$

Fix $\eta \in (0,1)$ (e.g., $\eta = \frac{1}{2}$) and set

$$\alpha \;:=\; (1+\eta)\frac{r}{d}.$$

Assume additionally that $r \le d/2$, so that $\alpha \in (0,1)$ and Lemma 8 applies. Then

$$s\left[1 - I_\alpha\!\left(\frac{r}{2}, \frac{d-r}{2}\right)\right] \;=\; s\,\Pr(B > \alpha) \;\le\; 2s\exp\!\left(-\frac{\eta^2 r}{72}\right), \tag{29}$$

where $B \sim \mathrm{Beta}\big(\frac{r}{2}, \frac{d-r}{2}\big)$.

We impose two conditions on r:

$$2s\exp\!\left(-\frac{\eta^2 r}{72}\right) \le \frac{1}{2}\,\delta_{\mathrm{Gauss}}(\varepsilon), \tag{30}$$

$$\alpha = (1+\eta)\frac{r}{d} \le \alpha_0. \tag{31}$$

Condition equation 31 is equivalent to $r \le \frac{\alpha_0}{1+\eta}\,d$, which is ensured whenever $r \ll d$. Condition equation 30 holds whenever

$$r \;\ge\; \frac{72}{\eta^2}\log\!\left(\frac{4s}{\delta_{\mathrm{Gauss}}(\varepsilon)}\right),$$

which is of the form $r \gtrsim \log s$ up to constant factors. Under these conditions we have, by monotonicity of $T(\varepsilon;\cdot)$ and equation 28,

$$T(\varepsilon;\alpha\mu) \;\le\; T(\varepsilon;\alpha_0\mu) \;=\; \frac{1}{2}\,\delta_{\mathrm{Gauss}}(\varepsilon),$$

and by equation 29 and equation 30,

$$s\left[1 - I_\alpha\!\left(\frac{r}{2}, \frac{d-r}{2}\right)\right] \;\le\; \frac{1}{2}\,\delta_{\mathrm{Gauss}}(\varepsilon).$$

Therefore,

$$\delta_{\mathrm{ours}}(\varepsilon;\alpha) \;\le\; T(\varepsilon;\alpha\mu) + s\left[1 - I_\alpha\!\left(\frac{r}{2}, \frac{d-r}{2}\right)\right] \;\le\; \delta_{\mathrm{Gauss}}(\varepsilon).$$

Moreover, the inequality is strict whenever the two half-budget bounds above are strict (e.g., by taking r slightly larger than the lower threshold and slightly smaller than the upper threshold), which yields

$$\delta_{\mathrm{ours}}(\varepsilon;\alpha) \;<\; \delta_{\mathrm{Gauss}}(\varepsilon).$$

Finally, by construction $\alpha = (1+\eta)\frac{r}{d}$, so $\alpha \approx r/d$, concluding the proof. □

F Comparison of Our Large-r Analysis with the Standard Analysis

We next compare the privacy guarantee in Theorem 3 to a classical baseline that ignores the randomness in M: it treats $V \mapsto MV$ as deterministic and applies the Gaussian mechanism calibrated to the Frobenius sensitivity.

Fix $V, V' \in \mathbb{R}^{d\times n}$. Let $J = J(V, V')$ with $k \coloneqq |J|$. Assume $\|v_j\|_2 = \|v'_j\|_2 = 1$ and $\|v_j - v'_j\|_2 \le \Delta_v$ for all $j \in J$, then

$$\|\Delta V\|_F^2 = \sum_{j\in J} \|v_j - v'_j\|_2^2 \le k\Delta_v^2. \tag{32}$$

A deterministic sensitivity bound gives $\|MV - MV'\|_F \le \|M\|\,\|\Delta V\|_F$. Using $\|M\| \approx \sigma_M^2(\sqrt{d} + \sqrt{r})^2$, we obtain

$$\varepsilon^{(k)}_{\text{class}} \;\le\; \sigma_M^2(\sqrt{d} + \sqrt{r})^2\,\sqrt{k}\,K_G, \tag{33}$$

where $K_G \coloneqq \frac{\Delta_v}{\sigma_G}\sqrt{2\ln\!\left(\frac{1.25}{\delta_G}\right)}$. By Theorem 3, the k-column guarantee composes as

$$\varepsilon^{(k)}_{\text{ours}} = k(\varepsilon_\parallel + \varepsilon_\perp),, \tag{34}$$

with $\varepsilon_\parallel, \varepsilon_\perp$ as specified in Lemmas 2 and 16.

Specifically, by Lemma 16 and setting $\sigma_M^2 = \frac{1}{r}$,

$$\varepsilon_\parallel = \frac{2\sigma_M^2}{\sigma_G}\left(\sqrt{d} + \sqrt{r} + \sqrt{2\ln\frac{3}{\delta_s}}\right)\left(\sqrt{2r(1-\beta_\perp^2)(1-\rho_\parallel) + 2p\beta_\perp^2(1-\rho_\perp)} + \sqrt{2\ln\frac{3}{\delta_s}}\Delta_v\right)\sqrt{2\ln\left(\frac{1}{\delta_\parallel}\right)}.$$

Here, $p \le s$ is the effective dimension of the revealed column-span seen through the random projection (formally $p = \operatorname{rank}(U^\top Z)$), and $s = \dim(S)$ with $S = \operatorname{span}(V_{-j})$.

Setting $\delta_\parallel = \delta_G/3$, and ignore all the small $\ln 1/\delta$ terms, we can derive an approximate upper bound on the ratio between our epsilon and the epsilon from the standard analysis (Equation (33)),

$$\begin{aligned}\frac{\varepsilon^{(k)}_{\text{ours}}}{\varepsilon^{(k)}_{\text{class}}} &\lessapprox \sqrt{k}\left(\underbrace{\frac{(\sqrt{d}+\sqrt{r})\sqrt{2r(1-\beta_\perp^2)(1-\rho_\parallel) + 2p\beta_\perp^2(1-\rho_\perp)}}{(\sqrt{d}+\sqrt{r})^2\Delta_v}} + \underbrace{\frac{\varepsilon_\perp}{\sigma_M^2(\sqrt{d}+\sqrt{r})^2\,K_G}}\right).\\ &= \underbrace{\frac{\sqrt{2kr(1-\beta_\perp^2)(1-\rho_\parallel) + 2kp\beta_\perp^2(1-\rho_\perp)}}{(\sqrt{d}+\sqrt{r})\Delta_v}}_{\text{Gaussian part}} + \underbrace{\frac{\sqrt{k}\varepsilon_\perp}{\sigma_M^2(\sqrt{d}+\sqrt{r})^2\,K_G}}_{\text{residual part}}.\end{aligned} \tag{35}$$

We can further bound $\varepsilon_\perp$ by Lemma 2,

$$\varepsilon_\perp \le \frac{\nu + \Delta_\perp}{2}\left(\frac{1}{\rho_\perp - K_\perp} + \log(\rho_\perp - K_\perp) - 1\right) - (r - 1 + {}^{\Delta_\perp}\!/_2)\log(\rho_\perp - K_\perp).$$

In particular, when $\rho_\perp$ is close to 1 and σ_G is small (hence K_G is large), the residual part is small, so the comparison is dominated by the Gaussian part.

G LoRA

LoRA adapts a pretrained weight matrix $W_0 \in \mathbb{R}^{n\times d}$ via a low-rank update $W_{\text{eff}} = W_0 + BA$, where $B \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times d}$ with $r \ll \min\{n, d\}$. In LoRA-FA, A is sampled once at initialization and then frozen, and only B is trained. Starting from B_0 (typically 0), each step uses $W = W_0 + B_{t-1}A$ on a minibatch $\mathcal{B}_t$ and updates $B_t = B_{t-1} - \eta\,\nabla_B\mathcal{L}(W; \mathcal{B}_t)$.

DP-LoRA-FA To train LoRA-FA on private data, it suffices to make the procedure that outputs B_t is DP, since A and W_0 are fixed (and not data-dependent). To the best of our knowledge, no DP-LoRA algorithm has been proposed for the fixed A setting, however to privatize LoRA-FA in the same spirit as Sun et al. (2024) we apply *per-example* gradient clipping to the gradients with respect to B and add Gaussian noise to the averaged clipped gradient. Concretely, for a minibatch $\mathcal{B}_t$ we compute per-example gradients $G_{t,i} = \nabla_B \ell(W; i)$, clip each to Frobenius norm at most β, $\widetilde{G}_{t,i} = \min\Big\{1, \frac{\beta}{\|G_{t,i}\|_F}\Big\} G_{t,i}$, and form the noisy DP gradient

$$\widehat{G}_t \;=\; \frac{1}{B_{\mathrm{mb}}} \sum_{i \in \mathcal{B}_t} \widetilde{G}_{t,i} \;+\; \frac{\sigma}{B_{\mathrm{mb}}} E_t, \qquad (E_t)_{jk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$$

which is then used to update B via $B_t = B_{t-1} - \eta\, \widehat{G}_t$. Because each $\widetilde{G}_{t,i}$ has Frobenius norm at most β, changing one example affects the summed clipped gradient by at most 2β in Frobenius norm. Adding Gaussian noise calibrated to β yields a differentially private update. Over T iterations (and with minibatch subsampling), the overall privacy guarantee for the final released B_T (and hence $W_0 + B_T A$) follows by standard DP-SGD privacy accounting/composition. Algorithm 1 describes T steps of one-layer DP-LoRA-FA.

Algorithm 1 DP-LoRA-FA (one layer, T steps)

Input: pretrained weight matrix $W_0 \in \mathbb{R}^{n \times d}$; rank $r \ll \min\{n, d\}$; steps T; dataset size N; minibatch size B_{mb}; per-example loss $\ell(\cdot;\cdot)$; step size η; clipping norm β; privacy parameters (ε, δ).

Set privacy noise $\sigma \leftarrow \frac{2\beta\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$.
Sample $A \sim \mathcal{N}(0, 1/r)^{r \times d}$ and freeze it.
Initialize $B \leftarrow 0 \in \mathbb{R}^{n \times r}$.
for $t = 1, \ldots, T$ **do**
  Sample minibatch $\mathcal{B}_t \subset [N]$ with $|\mathcal{B}_t| = B_{\mathrm{mb}}$ (e.g. Poisson rate $q = B_{\mathrm{mb}}/N$).
  Effective weights: $W \leftarrow W_0 + BA$.
  for each $i \in \mathcal{B}_t$ **do**
    $G_{t,i} \leftarrow \nabla_B \ell(W; i), \quad \widetilde{G}_{t,i} \leftarrow \min\Big(1, \frac{\beta}{\|G_{t,i}\|_F}\Big) G_{t,i}$.
  end for
  DP gradient: $\widehat{G}_t \leftarrow \frac{1}{B_{\mathrm{mb}}} \sum_{i \in \mathcal{B}_t} \widetilde{G}_{t,i} \;+\; \frac{\sigma}{B_{\mathrm{mb}}} E_t$, where $(E_t)_{jk} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.
  Update: $B \leftarrow B - \eta\, \widehat{G}_t$.
end for
Output: B (and implicitly $W = W_0 + BA$).

Multi-step Noisy projection mechanism with (M2) Algorithm 2 summarizes our algorithm, which applies M2 to SGD for a single layer over T optimization steps.

For $W = W_0 + BA$, the gradient with respect to B follows by the chain rule:

$$\nabla_B \mathcal{L}(W; \mathcal{B}_t) \;=\; \nabla_W \mathcal{L}(W; \mathcal{B}_t)\, A^\top,$$

Therefore, we can re-write the training dynamic above as, set $W = W_0 + B_0 A_0$ (where we assume B_0 was initialized as a 0 matrix) and update

$$B_1 = B_0 - \eta \nabla_W \mathcal{L}(W; \mathcal{B}_t)\, A^\top$$

further

$$W = W_0 + B_1 A = W_0 + B_0 A - \eta \nabla_W \mathcal{L}(W; \mathcal{B}_t)\, A^\top A = W_0 - \eta \nabla_W \mathcal{L}(W; \mathcal{B}_t)\, A^\top A$$

so if we wanted to privatize one single step of one-layer of LoRA-FA we would exclusively need to privatize

$$\nabla_W \mathcal{L}(W; \mathcal{B}_t)\, A^\top A$$

as W_0, B_0 are fixed. We can achieve this by clipping and adding noise:

$$\left(\min\left(1, \frac{\beta'}{\|\nabla_W \mathcal{L}(W; \mathcal{B}_t)\|_F}\right) \nabla_W \mathcal{L}(W; \mathcal{B}_t) + \sigma' E'\right) A^\top A$$

Algorithm 2 Noisy Projection Mechanism

Input: pretrained $W_0 \in \mathbb{R}^{n\times d}$; step size η; dataset size N; minibatch size B_{mb} (or full-batch); loss $\mathcal{L}$; clipping level β'; additive noise level σ'.

Notation: $\text{clip}_{\beta'}(X) = \min\left(1, \frac{\beta'}{\|X\|_F}\right) X$.

for $t = 1, \dots, T$ **do**
  Set $B \leftarrow 0 \in \mathbb{R}^{n\times r}$
  Sample $A \sim \mathcal{N}(0, 1/r)^{r\times d}$ Initialization
  Set $W_{t-1} \leftarrow W_{t-1} + BA$
  Sample minibatch $\mathcal{B}_t \subset [N]$ of size B_{mb}
  for each $i \in \mathcal{B}_t$ **do**
    $G_{t,i} \leftarrow \nabla_B \ell(W; i), \quad \widetilde{G}_{t,i} \leftarrow \min\left(1, \frac{\beta}{\|G_{t,i}\|_F}\right) G_{t,i}.$ Gradient Clipping
  end for
  $\widehat{G} \leftarrow \left(\frac{1}{B_{\text{mb}}} \sum_{i=1}^{B_{\text{mb}}} \tilde{G}_{t,i} + \sigma' E' A^\top\right) A$, where $E'_{ij} \sim \mathcal{N}(0,1)$ i.i.d. Compute DP gradient
  $W_t \leftarrow W_{t-1} - \eta \hat{G}$ Parameter Update
end for

Note that this is exactly the noisy projection mechanism studied in Section 4.2. We want to remark that W in the above equation does not depend on A as $B_0 = 0$, so we have that A is independent of the gradient and the privacy results of Section 4.2 are applicable.

G.1 Comparison between one step of DP-LoRA-FA and Projection Mechanism

In order to compare Algorithm 2 to Algorithm 1 we note that the final weights of DP-LoRA-FA (if we neglect clipping) are

$$\begin{aligned} W_T &= W_0 + BA \\ &= W_0 - \eta \left(\sum_{t=1}^{T} \nabla_B \mathcal{L} + \sigma E\right) A \end{aligned}$$

for our projection mechanism we have

$$\begin{aligned} W_T &= W_0 - \eta \left(\sum_{t=1}^{T} \nabla_W \mathcal{L} + \sigma' E'\right) A_t^\top A_t \\ &= W_0 - \eta \left(\sum_{t=1}^{T} \nabla_W \mathcal{L} A_t^\top + \sigma' E' A_t^\top\right) A_t \\ &= W_0 - \eta \left(\sum_{t=1}^{T} \nabla_B \mathcal{L} + \sigma' E' A_t^\top\right) A_t \end{aligned}$$

This means to compare of one step in one layer if we assume we set the clipping threshold in a way that we won't clip with high probability (we choose β and β' so that we do not clip most of the time) then we can simply compare the privacy of the releases

$$\nabla_B \mathcal{L} + \sigma \cdot E$$

to

$$\nabla_B \mathcal{L} + \sigma' E' A^\top.$$

Where we recall that $\sigma = \frac{2\beta\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$ and $\sigma' = \frac{2\beta'\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$. Further our analysis in Corollary 1 shows that the random projection $AA^\top$ will contract the sensitivity ΔV with a multiplier

of $\alpha \approx \frac{r}{d}$ w.h.p. for r small. This means the projection mechanism we implement to compare to DP-LoRA-FA actually has a noise multiplier of $\sqrt{r/d}\sigma'$. This means we are interested in comparing the following two mechanisms:

$$\mathcal{M}_{\text{DP-LoRA}}\left(\nabla_B \mathcal{L}\right) = \nabla_B \mathcal{L} + \sigma E$$
$$\mathcal{M}_{\text{ProjMech}}\left(\nabla_B \mathcal{L}\right) = \nabla_B \mathcal{L} + \sigma' \sqrt{\frac{r}{d}} E' A^\top$$

We would like to show that $\mathcal{M}_{\text{ProjMech}}$ (ε, δ)-DP $\implies$ $\mathcal{M}_{\text{DP-LoRA}}$ $(\tilde{\varepsilon}, \tilde{\delta})$-DP. For this we need to understand how the variances of σE and $\sigma' \sqrt{\frac{r}{d}} E' A^\top$ relate to each other. So first of all we will need to determine how σ and σ' differ. Notice they only differ in their clipping thresholds β and β'. We chose β in order to control the norm of $\nabla_B \mathcal{L}$ we choose β' in order to control the norm of $\nabla_W \mathcal{L}$. We note that $\nabla_W \mathcal{L} A^\top = \nabla_B \mathcal{L}$ by the chain rule, where $A \in \mathbb{R}^{r \times d}$ with coordinates sampled i.i.d. from $\mathcal{N}(0, 1/r)$. So we will use the Johnson-Lindenstrauss Lemma to compare the norms of $\nabla_W \mathcal{L} A^\top$ and $\nabla_W \mathcal{L}$ to find comparable choices of β and β'.

Lemma 19 (Johnson-Lindenstrauss (Dasgupta & Gupta, 2003)). *For A a random matrix $A \in \mathbb{R}^{r \times d}$ obtained from sampling the coordinates i.i.d from $\mathcal{N}(0, 1/r)$ and $x \in \mathbb{R}^d$ any non zero vector we have that*

$$(1 - \zeta)\|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \zeta)\|x\|_2^2$$

with probability $1 - 2e^{-\frac{r}{2}(\frac{1}{2}\zeta^2 - \frac{1}{3}\zeta^3)}$.

By the union bound, the probability that this relation is true for $x_1, \ldots, x_n$ is greater than $1 - 2ne^{-\frac{r}{2}(\frac{1}{2}\zeta^2 - \frac{1}{3}\zeta^3)}$

A common simplification is that for $\zeta \in (0, 1)$,

$$\frac{1}{2}\zeta^2 - \frac{1}{3}\zeta^3 \;\geq\; \frac{1}{6}\zeta^2,$$

since

$$\frac{1}{2}\zeta^2 - \frac{1}{3}\zeta^3 - \frac{1}{6}\zeta^2 = \frac{1}{3}\zeta^2(1 - \zeta) \;\geq\; 0.$$

Note $\|\nabla_W \mathcal{L} A^\top\|_F = \|A \nabla_W \mathcal{L}^\top\|_F$. So if we define $V = \nabla_W \mathcal{L}^\top$ we have that $V \in \mathbb{R}^{d \times n}$ and we can for ease of notation equivalently analyse $\|AV\|_F$ compared to $\|V\|_F$. For $V \in \mathbb{R}^{d \times n}$ it suffices to require

$$2n \exp\Big(-\frac{r}{12}\zeta^2\Big) \leq \delta_{\text{JL}}, \qquad \text{which implies} \qquad \zeta \;\geq\; \sqrt{\frac{12 \log(2n/\delta_{\text{JL}})}{r}}.$$

Hence, setting $\zeta = \sqrt{\frac{12 \log(2n/\delta_{\text{JL}})}{r}}$, with probability at least $1 - \delta_{\text{JL}}$,

$$\frac{\|AV\|_F}{\sqrt{1 + \zeta}} \;\leq\; \|V\|_F \;\leq\; \frac{\|AV\|_F}{\sqrt{1 - \zeta}}.$$

as $\|V\|_F^2 = \sum_j \|v_j\|_2^2$, for v_j the columns of V.

Recalling that β is the clipping bound for $\|AV\|_F = \|\nabla_W \mathcal{L} A^\top\|_F = \|\nabla_B \mathcal{L}\|_F$ and β' for $\|V\|_F = \|\nabla_W \mathcal{L}\|_F$ we have that whenever

$$\|\nabla_W \mathcal{L}\|_F \leq \beta' \implies \|\nabla_B \mathcal{L}\|_F \leq \sqrt{1 + \zeta}\beta'$$
$$\|\nabla_B \mathcal{L}\|_F \leq \beta \implies \|\nabla_W \mathcal{L}\|_F \leq \frac{1}{\sqrt{1 - \zeta}}\beta$$

For simplicity we set $\beta = \beta'$. This leads to $\sigma = \sigma'$. Next we want to compare the distribution of $\mathcal{M}_{\text{DP-LoRA}}$ and $\mathcal{M}_{\text{ProjMech}}$. Note they share the same mean ($\frac{\partial \mathcal{L}}{\partial B}$) so we are left to investigate their variance. For $\mathcal{M}_{\text{ProjMech}}$ the variance comes from $\sigma \sqrt{\frac{r}{d}} E' A^\top$ (recall we set $\beta = \beta'$, so $\sigma = \sigma'$).

Let

$$C \;\coloneqq\; \frac{r}{d} A A^\top \in \mathbb{R}^{r \times r}.$$

Since each row of $E' \in \mathbb{R}^{n\times d}$ is $\mathcal{N}(0, I_d)$, conditional on A we have

$$\sqrt{\frac{r}{d}}\, E' A^\top \,|\, A \;\sim\; \mathcal{MN}(0,\; I_n,\; C)\,,$$

equivalently (by row-stacking),

$$\mathrm{vec}\left(\sqrt{\frac{r}{d}}\, E' A^\top\right) \Big| A \;\sim\; \mathcal{N}(0,\; I_n \otimes C)\,.$$

Hence the mechanism

$$\mathcal{M}_{\text{ProjMech}}(D) \;=\; G(D) + \sigma\sqrt{\frac{r}{d}}\, E' A^\top$$

is (conditionally on A) a Gaussian mechanism with covariance

$$\Sigma_{\text{ProjMech}}(A) \;=\; \sigma^2\,(I_n \otimes C).$$

In comparison, the isotropic mechanism

$$\mathcal{M}_{\text{DP-LoRA}}(D) \;=\; G(D) + \sigma E, \qquad E \sim \mathcal{MN}(0, I_n, I_r),$$

has covariance

$$\Sigma_{\text{DP-LoRA}} \;=\; \sigma^2\,(I_n \otimes I_r).$$

To compare C to I_r note $A_{ij} \sim \mathcal{N}(0, 1/r)$ and write $A = \frac{1}{\sqrt{r}} Z^\top$ with $Z \in \mathbb{R}^{d\times r}$ i.i.d. $\mathcal{N}(0,1)$, so that

$$C = \frac{r}{d} AA^\top \;=\; \frac{1}{d} Z^\top Z.$$

were we recall $r < d$. We can use Lemma 7 to bound the eigenvalues of C and hence the amount of noise we add. Note that Lemma 7 is stated for $W = \frac{1}{r} ZZ^\top$, whose nonzero eigenvalues are of order d/r. In our setting we need bounds for $C = \frac{1}{d} Z^\top Z$, which has the same nonzero spectrum as $\frac{1}{d} ZZ^\top$. Since $\frac{1}{d} ZZ^\top = \frac{r}{d}\, W$, the eigenvalues rescale by the factor r/d, turning the $\sqrt{d/r}$ scale in Lemma 7 into a $\sqrt{r/d}$ deviation around 1. Consequently, for every $t \geq 0$, with probability at least $1 - 2e^{-t^2/2}$,

$$\left(1 - \sqrt{\frac{r}{d}} - \frac{t}{\sqrt{d}}\right)^2 I_r \;\preceq\; C \;\preceq\; \left(1 + \sqrt{\frac{r}{d}} + \frac{t}{\sqrt{d}}\right)^2 I_r.$$

Write $\varepsilon_{\text{proj}}(\alpha)$ and $\varepsilon_{\text{iso}}(\alpha)$ for the Rényi DP parameters of $\mathcal{M}_{\text{ProjMech}}$ and $\mathcal{M}_{\text{DP-LoRA}}$, respectively (of order $\alpha > 1$) at the same noise scale σ and the same clipping/sensitivity bound.

On the good event $\mathcal{E}$ where the spectrum of $C = \frac{r}{d} AA^\top$ concentrates, we have for all $\alpha > 1$,

$$(1-\eta)^2\, \varepsilon_{\text{proj}}(\alpha) \;\leq\; \varepsilon_{\text{iso}}(\alpha) \;\leq\; (1+\eta)^2\, \varepsilon_{\text{proj}}(\alpha), \qquad \eta \;\coloneqq\; \sqrt{\frac{r}{d}} + \sqrt{\frac{2\log(2/\gamma)}{d}}.$$

Consequently, after converting from RDP to (ε, δ)-DP, the isotropic mechanism satisfies

$$\mathcal{M}_{\text{ProjMech}} \;\; (\varepsilon_{\text{proj}}, \delta)\text{-DP} \quad\Longrightarrow\quad \mathcal{M}_{\text{DP-LoRA}} \;\; (\varepsilon_{\text{iso}}, \delta + \gamma)\text{-DP},$$

with ε_{iso} within a multiplicative factor $1 \pm O(\sqrt{r/d})$ of $\varepsilon_{\text{proj}}$ on $\mathcal{E}$ (assuming $d \gg r$ and $d \gg \log(1/\gamma)$).

H Additional Experimental Results

Additional MIA on noisy projection mechanisms (M1, M2) Additionally, we run noisy LoRA via gradient descent with $M(V + E)$, matching the small-r regime in Theorem 5. Results based on training 200 IN/OUT models with Equation (M2) are summarized in Table 2. As r increases, MIA success increases, consistent with Theorem 5. Additionally, Table 3 summarizes the results for MIA on Equation (M1) for larger r,

σ	Metric	Rank r 16	64	128	256	384	512
0.1	AUC	0.78	0.91	0.96	0.99	1.00	1.00
	Balanced Acc.	0.74	0.83	0.92	0.98	0.98	0.98
	Test Acc. (%)	44.61	49.98	54.52	57.94	60.38	61.70
0.5	AUC	0.56	0.70	0.74	0.76	0.82	0.86
	Balanced Acc.	0.56	0.67	0.70	0.71	0.75	0.80
	Test Acc. (%)	33.13	37.14	42.57	46.48	48.26	50.92

Table 2: MIA performance on (M2) across projection ranks r for two noise multiplier σ.

Metric	Rank r 32	63	128	512	800	1000
Avg. test acc. (%)	35.94	39.70	43.54	47.21	50.26	51.58
Balanced acc. (%)	61.50	65.00	67.50	72.00	72.00	81.00
AUC (%)	61.92	68.77	71.54	77.67	78.61	86.91

Table 3: MIA performance on (M1) projection rank r at noise multiplier 0.5 ($\varepsilon \approx 136.05$, $\delta = 10^{-5}$).

Empirical comparison of our accounting method and other baselines To empirically demonstrate the impact of our tighter privacy accounting, we compare the accuracy of our method (M2 with privacy accounting from Theorem 5; Algorithm 2) against DP-LoRA-FA, DP-LoRA, and DP-SGD under the same privacy budget on CIFAR-10 and SST-2. We fine-tune the linear head of a pretrained ResNet-50 on CIFAR-10 and RoBERTa-base on SST-2, with results summarized in Table 4. Table 5 includes test accuracy on CIFAR-10 for larger privacy paramters $\varepsilon \in \{0.5, 1, 2, 4\}$.

Our method achieves accuracy comparable to DP-LoRA and DP-LoRA-FA on both datasets, consistent with our analysis in Section 4.2 and Appendix G.1. Additionally, we observe that our method (and other low-rank approaches) outperforms DP-SGD on SST-2 but underperforms it on CIFAR-10.

Synthetic experiments for comparing our M2 and DP-LoRA-FA To isolate a regime in which resampling is more beneficial than DP-LoRA and DP-LoRA-FA as discussed in Section 4.2.2, we conduct a controlled multi-output linear-regression experiment with an approximately low-rank target and a non-negligible spectral tail. This setting favors our method because a fixed rank-r parameterization incurs persistent approximation bias by systematically omitting part of the spectral tail, whereas resampling the projection allows the optimization trajectory to accumulate updates across a broader collection of directions over time. Across 20 random seeds and both $\varepsilon \in \{4, 8\}$, our method consistently outperforms DP-LoRA and DP-LoRA-FA, with all 95% confidence intervals for the performance differences excluding zero.

Dataset	ε	DP-SGD	DP-LoRA	DP-LoRA-FA	Ours
CIFAR-10	0.15	0.81	0.76	0.79	0.8
	0.2	0.82	0.77	0.8	0.81
	0.3	0.83	0.79	0.81	0.82
SST-2	1.0	0.88	0.91	0.91	0.89
	2.0	0.88	0.91	0.91	0.9
	4.0	0.88	0.91	0.91	0.9

Table 4: Test accuracy comparison across different methods on CIFAR-10 and SST-2

Method	$\varepsilon = 0.5$	$\varepsilon = 1.0$	$\varepsilon = 2.0$	$\varepsilon = 4.0$
DP-SGD	–	–	84.75 ± 1.85	86.03 ± 1.17
DP-LoRA	88.86 ± 1.09	89.67 ± 0.88	89.74 ± 0.56	90.41 ± 0.68
Ours	84.72 ± 2.16	89.33 ± 0.00	86.31 ± 1.83	87.59 ± 1.46

Table 5: Test accuracy on CIFAR10 for larger ε

ε	**Comparator**	**Mean difference**	**95% CI**
4	DP-LoRA	$+0.004501$	$[+0.003664, +0.005338]$
	DP-LoRA-FA	$+0.019091$	$[+0.018378, +0.019803]$
8	DP-LoRA	$+0.014083$	$[+0.012967, +0.015198]$
	DP-LoRA-FA	$+0.057377$	$[+0.056517, +0.058237]$

Table 6: Comparison of our M2 against fixed-projection baselines on synthetic linear regression.

I Detailed setups of the experiments

In this section we discuss the experimental details. For our experiments on membership inference attack, each shadow model is trained on a single GPU, and for our experiments on comparing performance of M2, DP-LORA and DP-SGD, we run on CPU.

I.1 Membership inference attack

Datasets and pre-processing. Our target task is CIFAR-10. In the *pretrain+fine-tune* pipeline, we pretrain on CIFAR-100 and then fine-tune on CIFAR-10. All CIFAR inputs are normalized using per-channel mean $(0.4914, 0.4822, 0.4465)$ and standard deviation $(0.2023, 0.1994, 0.2010)$. During training we apply random crop (32×32 with padding 4) and random horizontal flip; during evaluation we apply normalization only.

Attack setup and notation. Let D be a CIFAR-10 subset of size $|D| = 5000$. The adversary selects a *canary* (x_q, y_q) and aims to infer whether it was included in training. Our membership inference evaluation procudure mimics the following membership inference game. We denote the (randomized) target-training procedure by $\mathcal{A}$ and the resulting trained model by f. For each attack trial, the model trainer samples

$$b \sim \text{Bernoulli}(0.5),$$

and trains the target model as

$$f_{\text{target}} \leftarrow \begin{cases} \mathcal{A}(D \cup \{(x_q, y_q)\}) & \text{if } b = 1 \quad \text{(IN)}, \\ \mathcal{A}(D) & \text{if } b = 0 \quad \text{(OUT)}. \end{cases}$$

Given f_{target}, D, (x_q, y_q), and knowledge of $\mathcal{A}$, the adversary outputs a guess $\hat{b} \in \{0, 1\}$.

Next, we instantiate our training algorithm, canary crafting algorithm, and membership inference evaluation method.

Training algorithm $\mathcal{A}$ (model, optimizer, and schedule). We use a CNN with three 3×3 convolution layers with channel sizes $32/64/128$, each followed by ReLU and 2×2 max-pooling, then two fully-connected layers. All training uses LoRA-FA gradient update based on SGD with momentum 0.9 and cross-entropy loss. Learning-rate schedules are either (i) cosine annealing over all iterations with minimum learning rate $\eta_{\min} = 10^{-4}$, or (ii) a step schedule.

LoRA-FA gradient updates. For each trainable weight tensor W in each layer, let G denote its minibatch gradient. We reshape G into a matrix $\tilde{G} \in \mathbb{R}^{d_o \times d_i}$ by flattening all non-output dimensions. At initialization, we sample a *fixed* Gaussian matrix $A \in \mathbb{R}^{r \times d_i}$ with i.i.d. entries $\mathcal{N}(0, 1)$ (one such

matrix per layer), and define

$$M \;=\; \frac{1}{r} A^\top A \in \mathbb{R}^{d_i \times d_i}.$$

We then apply a rank-r right-projection to the gradient,

$$\tilde{G} \;\leftarrow\; \tilde{G} M, \tag{36}$$

and reshape the projected gradient back to the original tensor shape before performing the optimizer update.

Canary construction. The adversary samples the canary input as $x_q \sim \mathcal{N}(0,1)$ with shape $3 \times 32 \times 32$. It then trains a reference model $f = \mathcal{A}(D)$ using the same procedure $\mathcal{A}$ as the model trainer, and sets the canary label to be the least-likely class under f:

$$y_q \;=\; \arg\min_{y \in \{0,\dots,9\}} [f(x_q)]_y,$$

where $f(x_q)$ denotes the logits and $[f(x_q)]_y$ is the logit for class y.

Membership inference evaluation protocol. To quantify membership leakage for the canary (x_q, y_q) under training algorithm $\mathcal{A}$, we train two collections of shadow models:

- IN models: $N_{\text{in}} = 1000$ models $\{f_i^{\text{in}}\}_{i=1}^{N_{\text{in}}}$ trained as $f_i^{\text{in}} \leftarrow \mathcal{A}(D \cup \{(x_q, y_q)\})$.
- OUT models $N_{\text{out}} = 1000$ models $\{f_j^{\text{out}}\}_{j=1}^{N_{\text{out}}}$ trained as $f_j^{\text{out}} \leftarrow \mathcal{A}(D)$.

For each trained model we compute the canary loss,

$$s_i^{\text{in}} = \ell\big(f_i^{\text{in}}, (x_q, y_q)\big)\,, \qquad s_j^{\text{out}} = \ell\big(f_j^{\text{out}}, (x_q, y_q)\big)\,.$$

We treat $\ell(\cdot)$ as a membership score: lower loss indicates higher likelihood of membership. We then estimate ROC-AUC and the best balanced accuracy directly from the two empirical score sets $\{s_i^{\text{in}}\}_{i=1}^{N_{\text{in}}}$ and $\{s_j^{\text{out}}\}_{j=1}^{N_{\text{out}}}$ by sweeping a threshold over all unique loss values.

Metrics. Given vectors of true memberships $\boldsymbol{b}$ and adversary predictions/scores $\hat{\boldsymbol{b}}$ (or scalar scores such as losses), we report: (i) ROC-AUC, and (ii) the best balanced accuracy obtained by sweeping thresholds over all unique scores, where balanced accuracy is $\frac{1}{2}(\text{TPR} + \text{TNR})$.

I.2 Experimental details for the comparison between our accounting method and other baselines

Setting for the experiments on CIFAR-10. We follow the representation-learning setup of Pinto et al. (2024) and use their fixed feature extractor[4]: a ResNet-50 pretrained with self-supervised learning on ImageNet-1K. For each CIFAR-10 example x_i, we compute a 2048-dimensional representation $z_i \in \mathbb{R}^{2048}$ (e.g., the pooled penultimate-layer feature), yielding a representation dataset $\{(z_i, y_i)\}_{i=1}^n$ with $z_i \in \mathbb{R}^{2048}$ and $y_i \in \{0, \dots, 9\}$. We then train a *linear* classifier on top of these frozen representations using three private training methods: (i) DP-SGD (Abadi et al., 2016), (ii) DP-LoRA-FA (Algorithm 1) and DP-LoRA, and (iii) our noisy-projection mechanism equation M2 (Algorithm 2) with the privacy accounting from Theorem 5. For method (iii), we resample an independent random projection matrix (equivalently, A_t and thus M_t) at every optimization step t. All the experiments on CIFAR-10 are conducted on CPUs, and the experiments on SST-2 are conducted on a single A100 GPU.

Setting for the experiments on SST-2. We fine-tune roberta-base on GLUE SST-2 using the standard single-sentence setup, using the codebase from Zhang et al. (2026)[5]. Each example is tokenized with the roberta-base tokenizer, truncated to a maximum length of 128, and dynamically padded within each batch, yielding a binary classification task with labels in $\{0, 1\}$; we train on the SST-2 training split and evaluate on the held-out validation split. Across all training algorithms, the

[4]Available at `https://github.com/FrancescoPinto/PILLAR`, licensed under CC BY 4.0.

[5]Available at `https://github.com/XidianNss/DP-SFT`, licensed under CC BY 4.0.

optimizer is AdamW with a linear warmup schedule with warmup ratio 0.1, and evaluation uses batch size 512. Similar to the CIFAR-10 experiments, we compare: i) DP-SGD implemented with AdamW under Opacus; ii) DP-LoRA, which inserts LoRA adapters into the query, value, key, and dense modules of RoBERTa; iii) DP-LoRA-FA, which additionally freezes the LoRA A matrices trains only the LoRA B matrices together with the classification head; and iv) our noisy-projection mechanism equation M2 (Algorithm 2) with the privacy accounting from Theorem 5.

Privacy accounting. Theorem 5 implies that a *single* (clipped) gradient update is $(\varepsilon,\ \delta_g + \delta_p)$-DP, where

$$\delta_g = T\left(\varepsilon, \frac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}\right) \qquad \text{and} \qquad \delta_p = s\left(1 - I_\alpha\left(\frac{r}{2}, \frac{d-r}{2}\right)\right).$$

Here, δ_g accounts for the privacy loss due to Gaussian noise addition (with noise level determined by σ_G), and δ_p upper-bounds the probability that the random projection fails to satisfy the required "good" event.

In our experiments we run compressed stochastic gradient descent for T steps, each step with a freshly sampled matrix M as in Equation (M2), and thus must compose privacy across steps. Although the mapping in equation M2 is not unconditionally a standard Gaussian mechanism, it becomes one after conditioning on

$$\mathcal{G}(M_t) \ := \ \left\{\frac{\|P_{M_t}\Delta V\|_F^2}{\|\Delta V\|_F^2} > \alpha\right\}.$$

Conditioned on $\mathcal{G}(M_t)$ (for fixed M_t), the update is equivalent to a Gaussian mechanism with effective ℓ_2-sensitivity $\sqrt{\alpha}\,\|\Delta V\|_F$ and Gaussian noise determined by Ξ. Therefore, conditioning on the intersection of "good" events across all steps, $\bigcap_{t=1}^{T}\mathcal{G}(M_t)$ (which are independent for freshly sampled M_t's), we can apply a standard tight Gaussian-mechanism accountant (e.g., via Rényi DP (Mironov, 2017)) to compose the Gaussian part over T steps (and incorporate privacy amplification by subsampling in the usual way). By a union bound, the intersection event holds with probability at least $1 - T\delta_p$, contributing an additive $T\delta_p$ term to the overall δ. Consequently, for fixed (δ_p, r, d) we compute the smallest admissible α, which yields an adapted effective sensitivity $\sqrt{\alpha}\,\|\Delta V\|_F$ for the Gaussian accountant. We then choose the noise multiplier to match the target privacy budget accordingly.

Hyperparameter selection. For experiments on CIFAR-10, we use batch size 1024 for all three methods. For DP-SGD, we use the learning rate recommended by Pinto et al. (2024) and tune the remaining hyperparameters by grid search over: number of epochs in $\{35, 40, 45, 50\}$ and clipping threshold in $\{0.5, 0.7, 1, 1.5, 2\}$. For DP-LoRA-FA, we tune over the same epoch and clipping grids, and additionally tune the learning rate in $\{0.1, 0.3, 0.5, 0.7, 1\}$ and the LoRA rank r in $\{32, 64, 128, 256, 512, 700\}$. For our mechanism, we set $\delta_p = 0.1\,\delta$ and optimize over the same hyperparameter grids as DP-LoRA-FA.

For experiments on SST-2, we fix the privacy parameter $\delta = 1e{-}5$ and select the best hyperparameters for each $\varepsilon \in \{0.5, 1, 2, 4\}$. As suggested by Zhang et al. (2026), we fix the LoRA alpha at 32, and tune the remaining hyperparameters by grid search over: number of epochs in $\{2, 3, 4\}$, clipping threshold in $\{0.5, 0.7, 1\}$, projection rank (or LoRA rank) in $\{16, 32, 64, 100, 128\}$, learning rate in $\{2e-4, 3e-4, 5e-4, 1e-3\}$, and weight decay in $\{0, 0.01\}$.

I.3 Data generating process for the large-r dataset in ??

Data generation. We take $d = 4096$ and $n = 64$ private columns, and split the coordinates into n disjoint blocks of size $k = d/n = 64$. For each group $j \in [n]$ we draw $N = 10^4$ raw records

$$x_{j,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}\big(c\,e_1,\ \sigma_{\text{data}}^2 I_k\big), \qquad i = 1, \ldots, N,$$

with public center norm $c = 1$ and $\sigma_{\text{data}} = 0.05$. Each record is clipped around the public center to radius $R = 0.5$:

$$\tilde{x}_{j,i} = c\,e_1 + (x_{j,i} - c\,e_1)\ \min\left\{1,\ \frac{R}{\|x_{j,i} - c\,e_1\|_2}\right\}.$$

We then average and normalize each group,

$$\bar{x}_j = \frac{1}{N}\sum_{i=1}^{N} \tilde{x}_{j,i}, \qquad u_j = \frac{\bar{x}_j}{\|\bar{x}_j\|_2} \in \mathbb{R}^k,$$

and embed u_j into the j-th coordinate block to form the query matrix $V \in \mathbb{R}^{d\times n}$:

$$V_{(j-1)k+1\,:\,jk,\;j} = u_j, \qquad \text{with all other entries of column } j \text{ equal to } 0.$$

Because the blocks are disjoint, $V^\top V = I_n$.

Neighboring datasets and sensitivity. Two datasets are neighbors if they differ by replacing a single raw record $x_{j,i}$ in one group j. Only column j of V can change. Since every clipped record satisfies $\|\tilde{x}_{j,i}\|_2 \geq c - R$, the sensitivity is bounded uniformly over all datasets by

$$\|v_j - v_j'\|_2 \;\leq\; D := \frac{2R}{N\,(c-R)} = 2\times 10^{-4}.$$

The remaining $s = n - 1$ columns stay unchanged and orthonormal, so $D_\| = 0$, $\beta_\perp = 1$, and $\rho_\perp \geq 1 - D^2/2$.

Mechanisms. Our large-r mechanism releases $MV + \Xi$, where $M = ZZ^\top$, $Z \in \mathbb{R}^{d\times r}$ has i.i.d. $\mathcal{N}(0, 1/r)$ entries, and $\Xi \in \mathbb{R}^{d\times n}$ has i.i.d. $\mathcal{N}(0, \sigma_G^2)$ entries. Both are independent of the data, and M is not released. The baseline is the Gaussian mechanism $V + \Xi$ with the same sensitivity D and the same noise level σ_G. We report (ε, δ) with $\delta = 10^{-5}$ and $\sigma_G \in \{10^{-3}, 3\times 10^{-3}\}$. All reported ε values are worst-case guarantees that hold uniformly over datasets. They depend only on $(d, n, r, D, \sigma_G, \delta)$, not on the particular sample drawn.