LoRA and Privacy: When Random Projections Help (and When They Don't)

Yaxi Hu*[1], Johanna Düngler*[2], Bernhard Schölkopf[1], and Amartya Sanyal[2]

[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany
{yaxi.hu, bernhard.schoelkopf}@tuebingen.mpg.de
[2]Department of Computer Science, University of Copenhagen
{jodu, amsa}@di.ku.dk
*Equal contribution.

Abstract

We introduce the *(Wishart) projection mechanism*, a randomized map of the form $S \mapsto Mf(S)$ with $M \sim \mathsf{W}_d(1/rI_d, r)$ and study its differential privacy properties. For vector-valued queries f, we prove non-asymptotic DP guarantees without any additive noise, showing that Wishart randomness alone can suffice. For matrix-valued queries, however, we establish a sharp negative result: in the noise-free setting, the mechanism is not DP, and we demonstrate its vulnerability by implementing a near perfect membership inference attack (AUC > 0.99). We then analyze a noisy variant and prove privacy amplification due to randomness and low rank projection, in both large- and small-rank regimes, yielding stronger privacy guarantees than additive noise alone. Finally, we show that LoRA-style updates are an instance of the matrix-valued mechanism, implying that LoRA is not inherently private despite its built-in randomness, but that low-rank fine-tuning can be more private than full fine-tuning at the same noise level. Preliminary experiments suggest that tighter accounting enables lower noise and improved accuracy in practice.

1 Introduction

Differential Privacy (DP) [Dwork et al., 2006] is widely regarded as the gold standard for protecting training data privacy in machine learning. Intuitively, DP limits the influence of any single example on the output, making it difficult to infer whether that example appeared in the training set. The most widely used DP algorithm in modern ML is DP-SGD [Abadi et al., 2016], the private counterpart of the de facto standard optimization algorithm Stochastic Gradient Descent (SGD).

However, DP-SGD is computationally demanding and often incurs a substantial utility loss, especially for large models. While it remains one of the few viable choices for training machine learning models from scratch, in many practical deployments sensitive data enters primarily during fine-tuning, e.g. when an organization adapts a public pre-trained model on proprietary data. This motivates a simple strategy: start from a large public pre-trained model and enforce privacy only during fine-tuning. In the non-private setting, parameter-efficient fine-tuning (PEFT) [Han et al.,

2024] updates only a small set of parameters while freezing the base model, substantially reducing memory and compute. This naturally raises the question: can PEFT similarly reduce the cost of DP fine-tuning?

Low-Rank Adaptation (LoRA) [Hu et al., 2022] is a widely used PEFT method that often matches full-parameter fine-tuning on downstream tasks [Dettmers et al., 2023, Hu et al., 2022]. It freezes the pre-trained weights and inserts randomly initialised trainable low-rank matrices, dramatically shrinking the number of trainable parameters. Variants include adaptive-rank methods [Zhang et al., 2023b], quantization-aware tuning for low-bit backbones [Dettmers et al., 2023, Li et al., 2024], stability/initialization refinements [Hayou et al., 2024, Meng et al., 2024], and structural decompositions [Liu et al., 2024], each targeting better accuracy under tight compute and memory budgets. Approaches to privatising LoRA have also been proposed, including DP-LoRA [Liu et al., 2025].

Several LoRA variants [Hao et al., 2024, Sun et al., 2024] already incorporate substantial randomness (e.g., repeated re-initialization of component weight matrices). Yet most privatisation algorithms treat LoRA updates as deterministic given the data and do not explicitly leverage this inherent randomness. At the same time, empirical studies report reduced memorisation under LoRA [Hong et al., 2025] even without any explicit privatisation.

Following Hao et al. [2024], we note that in certain LoRA variants (e.g. LoRA-FA [Zhang et al., 2023a], the update behaves as if it applies a random Wishart projection to the gradient. We formalize this behaviour through the following abstraction, (which allows us to rigorously analyse the privacy guarantee of these methods)

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]*.*

We therefore analyse $\mathcal{A}_r$ as a proxy for LoRA's built-in randomness. Unlike the Johnson–Lindenstrauss [Johnson and Lindenstrauss, 1984] transform, which is a random embedding that preserves norms, the Wishart projection is a random positive semidefinite reweighting that is isotropic in expectation but can distort norms in a single draw.

Recent work [Malekmohammadi and Farnadi, 2024] conjectures that LoRA training dynamics may inherently possess privacy properties. Specifically, they show that individual rows of the gradient matrix asymptotically resembles the noise distribution of DP-SGD as the dimension $d \to \infty$. In contrast, we establish non-asymptotic differential privacy guarantees for the vector-valued projection mechanism in Theorem 1. However, LoRA updates are matrix-valued: in our abstraction, they correspond to applying the projection mechanism to gradient matrices. We show that this matrix-valued projection mechanism, and hence LoRA, is not private in general (Proposition 2): in the setting without additional additive noise, neighbouring datasets can yield non-overlapping support, enabling (near) perfect membership inference with AUC ≥ 0.99 as illustrated in Table 1.

[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).

Nevertheless, LoRA's built-in randomness is not irrelevant: it is simply insufficient on its own. We therefore consider a noisy variant of the projection mechanism, obtained by adding a small amount of additive noise. This enforces overlap for 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 projection 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.

We summarize our contributions below.

(i) We introduce the Wishart projection mechanism and prove non-asymptotic DP guarantees in the vector case.

(ii) We show the corresponding matrix mechanism is not DP, and demonstrate its vulnerability via a near-perfect MIA (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 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

We first review DP preliminaries, then introduce LoRA and show how it instantiates the projection mechanism.

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,

$$\Delta \coloneqq \max_{S \sim_H S'} \left\| f(S) - f(S') \right\|_2 .$$

A canonical example is the Gaussian mechanism:

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 and 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., 2023a] is a variant that fixes A and only updates B. When A is initialized Gaussian and then *frozen*, we update B by

$$B_{t+1} \ = \ B_t - \eta\left(\nabla_W \mathcal{L}(W_t)\right)A^\top.$$

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 and 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 in Section 3.1 that it is differentially private under an alignment assumption. In contrast, for matrix-valued queries we prove a negative result: Proposition 2 shows the privacy loss is unbounded ($\varepsilon = \infty$ for any $\delta < 1$). This also implies that LoRA-FA and LoRA is not private.

Setting	Metric	Rank r				
		16	**32**	**128**	**256**	**512**
Train from scratch	Test Acc (%)	57.31	61.76	66.36	67.34	68.36
	AUC (%)	99.86	99.64	100.00	100.00	100.00
Pretrain+Finetune	Test Acc (%)	75.98	75.63	76.12	76.04	76.45
	AUC (%)	99.81	99.98	100.00	100.00	100.00

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

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 $(\varepsilon_\rho, \delta_\rho)$ 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 B.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 $\mathbb{R}^d$, let $\rho > 0$ be the minimum alignment for f on $\mathcal{D}$ as defined in Equation (2). Let $\delta' > 0$. If $\rho > \frac{t_r(1-\delta')}{\sqrt{r+t_r(1-\delta')^2}}$, then, the projection mechanism (Definition 1) with rank r is $(\varepsilon_\rho, \delta_\rho)$-DP on $\mathcal{D}$, with*

$$\delta_\rho = \mathbb{E}_{x \sim \chi^2_r}\left[\Phi\left(-\frac{\rho\sqrt{x}}{\sqrt{1-\rho^2}}\right)\right] + 3\delta'$$

$$\varepsilon_\rho \le \frac{d-r+1}{2}\ln(\rho + K) + \frac{(1-\rho+K)\kappa_{d+r-1}(1-\delta')}{2(\rho - K)}$$

where $K = \sqrt{\frac{1-\rho^2}{r}}t_r(1-\delta')$.

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, for fixed δ the resulting ε depends non-monotonically on r. This reflects 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 relies on fewer Gaussian directions, increasing tail risk and potentially weakening the (ε, δ) guarantee.

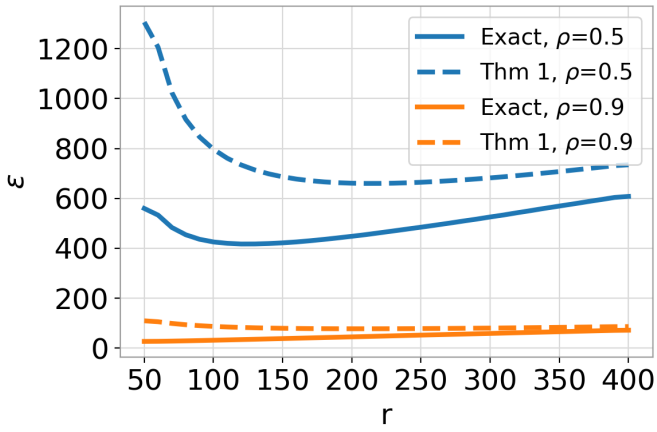

Figure 1: Privacy loss ε v.s. rank r at $\delta = 0.01$ and $d = 400$: Monte Carlo simulation of exact privacy profile and Theorem 1.

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 $f(S) \coloneqq \tilde{f}(S)/\|\tilde{f}(S)\|_2$ 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 $\|_2\nabla\ell(\theta; x)\|_2 \le L$ and $\|\tilde{f}(B)\|_2$ is bounded away from zero, then $\rho = 1 - \widetilde{O}(1/|B|^2)$.

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 a 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 with*

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

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

Proof intuition: Let $V \coloneqq f(S)$ and $V' \coloneqq f(S')$. As $V \neq V'$ such that $V - V' \neq 0$ and $\operatorname{rank}(V - V') = s \ge 1$. Given S the mechanism outputs the random variable $Y = VM = VZZ^\top$, where each column z_i of Z is i.i.d. $\mathcal{N}(0, I_d/r)$. For the two outputs to coincide we require $(V - V')M = 0$, i.e. $(V - V')Z = 0$: every random direction z_i must be orthogonal to the rowspace of $(V - V')$. As

$\text{rank}(V - V') = s$, its orthogonal complement is a $(d-s)$-dim subspace. A continuous Gaussian vector z_i lands in a fixed lower dimensional subspace with probability 0. Hence, $VM \neq V'M$ almost surely. Taking $E = \text{Supp}(VM)$ concludes the proof. (See Lemma 11).

Proposition 2 extends beyond our abstract projection mechanism to directly rule out "privacy for free" arising from LoRA's random initialization. Specifically, the first update step of standard LoRA (where $B_0 = 0$) is equivalent to the Definition 1, implying standard LoRA cannot satisfy DP without additional additive noise (See Appendix C.1 for detailed explanation).

Motivated by the negative result above, we test noise-free LoRA against a membership inference attack (MIA). 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 test accuracy and ROC-AUC. 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 F.1 for details.

These findings do not conflict with prior work showing that LoRA can reduce MIA success [Hong et al., 2025, Malekmohammadi and Farnadi, 2024], 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) given by random subspace hiding.

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]*:*

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

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

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

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, 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 $n-1$ columns, an adversary can learn information about M and use it to infer the remaining column more accurately. Thus additive noise Ξ is needed to mask this leakage and ensure distributional overlap.

4.1.1 Setting and Main Theorem

We first prove a privacy guarantee for a one-column adjacency, and then lift it to a more general matrix adjacency by composing such one-columns steps. Before stating our results, we define these adjacency relations in Definition 4.

Throughout this section we assume all columns are unit vectors: $\|v_j\|_2 = \|v'_j\|_2 = 1$. For $V, V' \in \mathbb{R}^{d\times n}$, write $J(V, V') \coloneqq \{j \in [n] : v_j \neq v'_j\}$, and for any $j \in [n]$, let V_{-j} denote V with column j removed.

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 $S \coloneqq \operatorname{span}(V_{-j}) \subset \mathbb{R}^d$. We write $V \sim_A^{(j)} V'$ 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*

$$\|b\|_2, \|b'\|_2 = \beta_\perp, \ \frac{\langle a, a'\rangle}{\|a\|_2 \|a'\|_2} \geq \rho_\parallel, \ \frac{\langle b, b'\rangle}{\|b\|_2 \|b'\|_2} \geq \rho_\perp.$$

2. ***(Matrix adjacency).*** *We write $V \sim_A V'$ if there exists an ordering $j_1, \ldots, j_k$ of $J(V, V')$ where $k = |J(V, V')|$, matrices $V^{(0)}, \ldots, V^{(k)}$ such that $V^{(0)} = V$, $V^{(k)} = V'$, and $V^{(t-1)} \sim_A^{(j_t)} V^{(t)}$ for all $t \in [k]$.*

The parameter $\beta_\perp$ measures how much of the changed column lies outside the span exposed by the other 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 $V \sim_A V'$ with $k = |J(V, V')|$. Define $(\varepsilon_{\text{one}}, \delta_{\text{one}})$ as the privacy parameters for a single admissible step $V \sim_A^{(j)} V'$ (given in Theorem 4). Then $\mathcal{A}_1$ is (ε, δ)-DP for inputs V, V', with*

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

4.1.2 Proof Sketch

Let $V^{(0)}, \ldots, V^{(k)}$ 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 $V \sim_A^{(j)} V'$. The matrices agree on all columns except j so $V_{-j} = V'_{-j}$. Write the mechanism's output as:

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

Let $S = \mathrm{span}(V_{-j})$ with orthonormal basis U, and set $G \coloneqq U^\top Z$. Using the decomposition $M = M_\parallel + M_\perp$ from Appendix D.1, we have: (i) $M_\perp V_{-j} = 0$, so Y depends only on $M_\parallel$; and (ii) *posterior stability*: conditional on G, the residual block $M_\perp$ remains independent of Y. This is the key decoupling: after observing the unchanged columns, the randomness in $M_\perp$ is still "fresh".

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

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

We bound these two parts separately and then compose. **Residual term.** For the residual term $R(b)$, posterior stability (Lemma 13) lets us apply the *vector-valued* DP analysis from Section 3, yielding an $(\varepsilon_\perp, \delta_\perp)$ bound. Let Φ be the standard normal CDF; let t_q and κ_ℓ be the quantiles of Student t and χ^2_ℓ distributions, respectively.

Lemma 2 (Residual DP bound)**.** *Write $s \coloneqq \dim(S)$ and $m \coloneqq d - s$. Fix $\rho_\perp \in (0, 1]$ and $\delta' \in (0, 1)$, and define*

$$K_\perp = \sqrt{\tfrac{1-\rho_\perp^2}{q}}\, t_q\left(1 - \delta'/3\right),$$

Then the residual release $R(b) = M_\perp b$ is $(\varepsilon_\perp, \delta_\perp)$-DP for inputs b, b' satisfying Definition 4 where

$$\delta_\perp = \mathbb{E}_{X \sim \chi^2_q}\left[\Phi\left(-\tfrac{\rho_\perp \sqrt{X}}{\sqrt{1-\rho_\perp^2}}\right)\right] + \delta',$$

$$\varepsilon_\perp \le \tfrac{m-q+1}{2} \ln(\rho_\perp + K_\perp) + \frac{\kappa_{m+q-1}\left(1 - \delta'/3\right)}{2}\left(\tfrac{1}{\rho_\perp - K_\perp} - 1\right).$$

Correlated term. For the correlated term $C(v_j)$, we apply the Gaussian mechanism with a high-probability directional sensitivity bound for $M_\parallel$: with probability at least $1 - \delta_s$ over Z, $\|M_\parallel(v_j - v'_j)\|_2 \le \Gamma_{\delta_s}\|v_j - v'_j\|_2$ (Lemma 14). Conditioning on this event we get the correlated term is $(\varepsilon_\parallel, \delta_\parallel) - DP$ with

$$\varepsilon_\parallel \coloneqq \frac{\Gamma_{\delta_s}\Delta_v}{\sigma_G}\sqrt{2\ln\left(1.25/\delta_\parallel\right)}. \tag{3}$$

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 $V \sim_A^{(j)} V'$ 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 $p = \mathrm{rank}(U^\top Z)$ and define Γ_{δ_s} as above, and $\varepsilon_\parallel$ as in Equation (3). 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*

$$\varepsilon_{\text{one}} \le \varepsilon_\parallel + \varepsilon_\perp, \qquad \delta_{\text{one}} \le \delta_\parallel + \delta_\perp + \delta_s.$$

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

4.1.3 Comparison with the Gaussian Mechanism

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 $k := |J|$. Assume $\|v_j\|_2 = \|v'_j\|_2 = 1$ and $\|v_j - v'_j\|_2 \le \Delta_v$ for all $j \in J$, then

$$\|\Delta V\|_F^2 = \sum_{j\in J} \|v_j - v'_j\|_2^2 \le k\Delta_v^2. \tag{4}$$

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

$$\varepsilon_{\text{class}}^{(k)} \;\le\; \sigma_M^2(\sqrt{d} + \sqrt{r})^2\,\sqrt{k}\,K_G, \tag{5}$$

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

$$\varepsilon_{\text{ours}}^{(k)} = k(\varepsilon_\parallel + \varepsilon_\perp),\ \text{with } \varepsilon_\parallel \le \sigma_M^2(\sqrt{d} + \sqrt{p})\sqrt{p}\,K_G, \tag{6}$$

where $p \le s$ 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 $S = \operatorname{span}(V_{-j})$.

For comparison, we set $\delta_\parallel = \delta_G/3$. Combining Equation (5) and Equation (6) yields

$$\frac{\varepsilon_{\text{ours}}^{(k)}}{\varepsilon_{\text{class}}^{(k)}} \;\le\; \sqrt{k}\left(\underbrace{\frac{(\sqrt{d} + \sqrt{p})\sqrt{p}}{(\sqrt{d} + \sqrt{r})^2}}_{\text{Gaussian part}} + \underbrace{\frac{\varepsilon_\perp}{\sigma_M^2(\sqrt{d} + \sqrt{r})^2\,K_G}}_{\text{residual part}}\right). \tag{7}$$

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

$$\varepsilon_\perp \le \underbrace{\frac{m-q+1}{2}\ln(\rho_\perp + \alpha)}_{T_1} \;+\; \underbrace{\frac{K}{2}\left(\frac{1}{\rho_\perp - \alpha} - 1\right)}_{T_2}.$$

where $m = d - s$, $q = r - p$ and α and K are the $(1 - \delta'/3)$-quantiles defined in Lemma 2.

In particular, when $\rho_\perp$ is close to 1 and α is small (e.g., for larger q and moderate δ'), the residual part is small, so the comparison is dominated by the Gaussian part. In the rank-deficient regime $s < r$, we moreover have $p = s$ a.s., so the Gaussian coefficient simplifies to

$$\frac{(\sqrt{d} + \sqrt{s})\sqrt{s}}{(\sqrt{d} + \sqrt{r})^2}. \tag{8}$$

In this regime, the coefficient is < 1, so our bound improves on the classical baseline.

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 $\mathcal{A}_2$ 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. In rare alignment events where $\Delta V \approx P_M \Delta V$ the amplification disappears and the guarantee becomes comparable to the naive approach.

4.2.1 Main Theorem

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

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

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 $\varepsilon > 0$, the mechanism $\mathcal{A}_2$ is $(\varepsilon, \delta_\alpha(\varepsilon))$-DP with*

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

where $I_\alpha(a, b)$ is the regularized incomplete Beta function (i.e. $I_\alpha(a, b) = \Pr[B \leq \alpha]$ for $B \sim \mathrm{Beta}(a, b)$).

4.2.2 Proof Sketch

Fix V, V' and condition on M. Let P_M denote the orthogonal projector onto $\operatorname{col}(M)$. Conditioned on M, the output $M(V + E)$ is a Gaussian matrix but supported on the rank-r subspace $\operatorname{col}(M)$. Consequently, the privacy loss depends on ΔV only through its projection onto $\operatorname{col}(M)$, namely $\|P_M \Delta V\|_F$. Let $P(\cdot \mid M)$ and $Q(\cdot \mid M)$ be the conditional output laws under inputs V and V', respectively. Since $\sigma_G > 0$ we have $P(\cdot \mid M) \ll Q(\cdot \mid M)$, so we can define the conditional privacy loss random variable

$$L_M(Y) \;\coloneqq\; \log \frac{dP(\cdot \mid M)}{dQ(\cdot \mid M)}(Y), \quad Y \sim P(\cdot \mid M).$$

Then as shown in Lemma 16 we have,

$$L_M(Y) \sim \mathcal{N}\left(-\tfrac{\|P_M \Delta V\|_F^2}{2\sigma_G^2}, \tfrac{\|P_M \Delta V\|_F^2}{\sigma_G^2}\right).$$

This motivates the following good-projection event

$$\mathcal{G}_\alpha \coloneqq \left\{M : \|P_M \Delta V\|_F^2 \leq \alpha \|\Delta V\|_F^2\right\}, \quad 0 < \alpha < 1.$$

On the good event $\mathcal{G}_\alpha$, $L_M(Y)$ has a Gaussian tail:

$$\Pr\left(|L_M(Y)| > \varepsilon \mid M\right) \le T\left(\varepsilon; \tfrac{\alpha\|\Delta V\|_F^2}{\sigma_G^2}\right). \tag{9}$$

Thus the unconditional δ is the Gaussian tail on $\mathcal{G}_\alpha$ plus the failure probability $\Pr(\mathcal{G}_\alpha^c)$. It remains to bound $\Pr(\mathcal{G}_\alpha^c)$.

Let $\operatorname{rank}(\Delta V) = s$, and let $u_1, \dots, u_s \in \mathbb{R}^d$ be an orthonormal basis for $\operatorname{col}(\Delta V)$. If $\|P_M u_i\|_2^2 \le \alpha$ for all $i \in [s]$, then $\|P_M \Delta V\|_F^2 \le \alpha \|\Delta V\|_F^2$. Thus, the bad event $\mathcal{G}_\alpha^c$ implies $\|P_M u_i\|_2^2 > \alpha$ for some i. Hence, by a union bound,

$$\Pr(\mathcal{G}_\alpha^c) \le \sum_{i=1}^{s} \Pr\left(\|P_M u_i\|_2^2 > \alpha\right).$$

Since $M = ZZ^\top$ with $Z \in \mathbb{R}^{d\times r}$ Gaussian, $\operatorname{col}(M) = \operatorname{col}(Z)$ is a random Haar rank-r subspace of $\mathbb{R}^d$, and P_M is the orthogonal projector onto this subspace (Lemma 17). For a fixed unit vector $u \in \mathbb{R}^d$, the captured energy $\|P_M u\|_2^2$ then follows the Beta law $\mathrm{Beta}\left(\frac{r}{2}, \frac{d-r}{2}\right)$. Therefore,

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

Combining this with equation 9 yields Theorem 5.

4.2.3 Comparison with the Gaussian Mechanism

Without conditioning on M, the standard Gaussian analysis uses sensitivity $\|\Delta V\|_F$ and gives, for any $\varepsilon > 0$,

$$\delta_{\mathrm{Gauss}}(\varepsilon) \;=\; T\left(\varepsilon; \frac{\|\Delta V\|_F^2}{\sigma^2}\right). \tag{10}$$

This is robust but pessimistic, as it treats the full gap $\|\Delta V\|_F$ as visible. Conditioning on M yields a tighter bound: Theorem 5 splits δ into a Gaussian term on the good event $\mathcal{G}_\alpha$ and a failure probability:

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

On $\mathcal{G}_\alpha$, the effective squared sensitivity drops from $\|\Delta V\|_F^2$ to $\alpha\|\Delta V\|_F^2$, so the Gaussian term improves as α decreases. Meanwhile $\Pr(\mathcal{G}_\alpha^c)$ decreases as α increases. Thus α trades off a tighter Gaussian term against a rarer good event. This trade-off translates into a concrete improvement once we optimize over α via simple numerical methods. As shown in Figure 2, the resulting bound can strictly beat the Gaussian baseline in the small-r regime (with larger gains for smaller r). Intuitively, a random rank-r subspace captures only $\approx r/d$ of the energy of a fixed direction. In the small-r regime, choosing α just above this typical level keeps the failure probability small while still substantially tightening the Gaussian term. Corollary 1 in the Appendix makes this precise: if $r \ll d$ and $\log s \lesssim r$, an appropriate α always outperforms the Gaussian baseline.

To relate these bounds to common training procedures, note that the Gaussian baseline above is exactly the privacy accounting for standard DP-SGD. Recalling that by our notation $V = \nabla_W \mathcal{L}(W_t)^\top$, privatizing V directly (without conditioning on M) amounts to adding Gaussian noise to the full

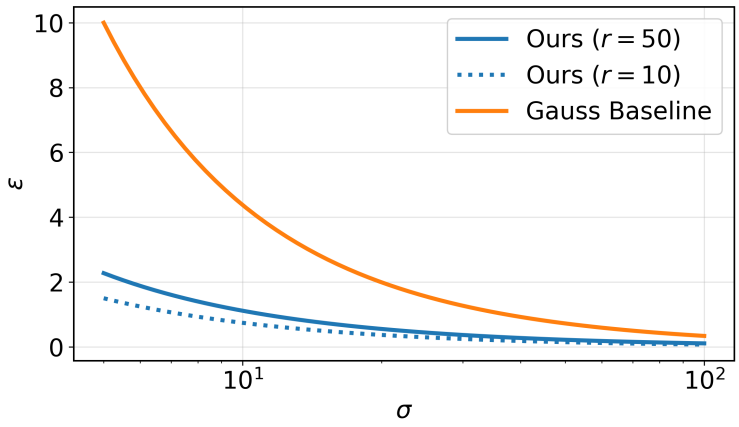

Figure 2: Privacy loss ε v.s. noise scale σ for $\delta = 1e-5$ and $d = 2000$: We compare our bound (Theorem 5), minimized over α with the Classical Gaussian mechanism (Equation (10)).

gradient, i.e., updating without any low-rank projection. We use $\mathcal{A}_2$ as a proxy for DP-LoRA-FA, but emphasize that it is not identical to the standard implementation: we add Gaussian noise to $V^\top = \nabla_W \mathcal{L}(W_t)$ and then apply the random projection, whereas DP-LoRA-FA adds noise after first multiplying by $A^\top$. We focus on the former variant because it admits a more tractable theoretical analysis. Nonetheless, Appendix E.1 shows that when analysing one training step, the two procedures inject noise of the same order, with high probability. For completeness, we include the standard DP-LoRA-FA implementation in our experiments.

For an empirical comparison, we fine-tune the linear head of a pretrained ResNet-50 on CIFAR-10 and evaluate our method (M2 with privacy accounting from Theorem 5) against DP-LoRA-FA and DP-SGD under the same privacy budget. Appendix F.2 provides implementation details (backbone, hyper-parameter grid, and accounting).

Figure 3 (right) plots test accuracy versus the privacy budget ε for $\delta = 10^{-4}$. In the small-ε regime, our method outperforms DP-SGD because the low-rank projection reduces the noise required to attain a fixed privacy level. DP-LoRA-FA shows the same effect, but only at even smaller ε. As ε grows, less noise is required noise for all methods and this noise advantage fades. For large ε the accuracy gains from reduced noise are outweighed by the optimization bias introduced by the low-rank constraint, so DP-SGD matches or slightly outperforms our approach. In Figure 3 (left), we isolate the effect of the low-rank projection by comparing our method with standard DP-LoRA-FA for varying ranks r, at fixed privacy budget ($\varepsilon \in \{0.2, 0.4\}$)[3]. While the two approaches are similar and, with high probability, add effective noise of the same order (see Appendix E.1), we observe that our Wishart projection consistently performs better, particularly at smaller ranks. This suggests that utility is influenced not only by the noise level, but also by the geometry of the random projection (applied to both the gradient matrix and the noise). Understanding when and why different projection distributions yield better utility is an interesting direction for future work.

[3]See results for more ε in Figure 4.

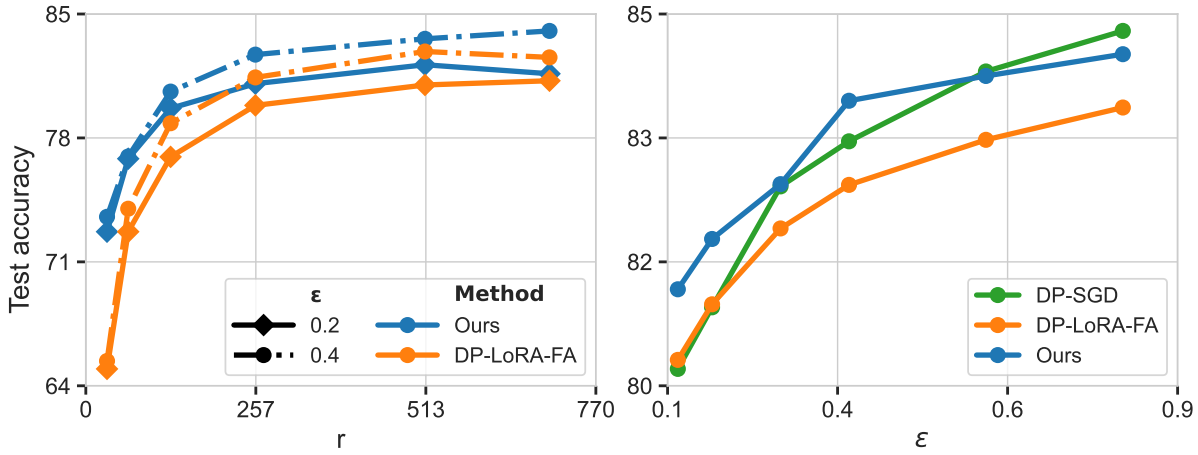

Figure 3: Comparison of DP-SGD, DP-LoRA-FA, and our mechanism (M2) using the privacy accounting of Theorem 5.

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 and 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 and 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 (resampled) low-dimensional random projections, but their motivation is primarily *efficiency*, not privacy. Malekmohammadi and Farnadi [2024] argue via a CLT-based approximation that LoRA can induce DP-SGD-like training dynamics for certain architectures in the limit. In contrast, we provide *non-asymptotic* results for the finite-dimensional updates used in LoRA: we give formal differential privacy guarantees for vector valued updates, a sharp non-privacy result for noise-free

matrix updates, and finally show that when adding small amounts of additive noise to the gradients the 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 $r \geq 128$, 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.

6 Acknowledgement

JD acknowledges support from the Danish Data Science Academy, which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516). AS acknowledges the Novo Nordisk Foundation for support via the Startup grant (NNF24OC0087820) and VILLUM FONDEN via the Young Investigator program (72069). The authors would also like to thank Rasmus Pagh for very insightful discussions.

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.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

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.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856.

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

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, 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.

William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. *Contemporary Mathematics*, 1984.

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, pages 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*, pages 177–182, 2011.

Yixiao Li, Yifan Yu, Chen Liang, Pengcheng He, Nikos Karampatziakis, Weizhu Chen, and Tuo Zhao. Loftq: Lora-fine-tuning-aware quantization for large language models. In *International Conference on Learning Representations*, 2024.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Forty-first International Conference on Machine Learning*, 2024.

Xiao-Yang Liu, Rongyi Zhu, Daochen Zha, Jiechao Gao, Shan Zhong, Matt White, and Meikang Qiu. Differentially private low-rank adaptation of large language model using federated learning. *ACM Trans. Manage. Inf. Syst.*, 2025.

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

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 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, editors, *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.

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*, 2023a.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023b.

Appendix

A Mathematical Preliminaries

We recall a few standard distributional facts used throughout the paper.

Lemma 3 (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 4 (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 5. *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 6 (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*

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

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

$$\left(\sqrt{\frac{d}{r}} - 1 - \frac{t}{\sqrt{r}}\right)^2 \;\leq\; \lambda^+_{\min}(W) \;\leq\; \lambda_{\max}(W) \;\leq\; \left(\sqrt{\frac{d}{r}} + 1 + \frac{t}{\sqrt{r}}\right)^2, \tag{11}$$

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 11. 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 7 (Tail bound for random capture fraction)**.** *Let $r \leq d/2$ and let*

$$B \sim \operatorname{Beta}\left(\frac{r}{2}, \frac{d-r}{2}\right).$$

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:

$$\{B > \alpha\} \;\subseteq\; \Big\{X > (1+\eta/3)r\Big\} \;\cup\; \Big\{Y < (1-\eta/3)(d-r)\Big\}. \tag{12}$$

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

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

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

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

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 12.

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

$$\Pr\big(Z > (1+t)k\big) \le \exp\Big(-\frac{t^2 k}{8}\Big), \qquad \Pr\big(Z < (1-t)k\big) \le \exp\Big(-\frac{t^2 k}{8}\Big). \tag{13}$$

Using equation 12 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 13 yields

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

and similarly

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

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

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

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

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

B Privacy Analysis of Vector Projection Mechanism

Theorem 1. *For a dataset collection $\mathcal{D}$ and a query function f with outputs in $\mathbb{R}^d$, let $\rho > 0$ be the minimum alignment for f on $\mathcal{D}$ as defined in Equation (2). Let $\delta' > 0$. If $\rho > \frac{t_r(1-\delta')}{\sqrt{r+t_r(1-\delta')^2}}$, then, the projection mechanism (Definition 1) with rank r is $(\varepsilon_\rho, \delta_\rho)$-DP on $\mathcal{D}$, with*

$$\delta_\rho = \mathbb{E}_{x\sim\mathcal{X}_r^2}\left[\Phi\left(-\frac{\rho\sqrt{x}}{\sqrt{1-\rho^2}}\right)\right] + 3\delta'$$

$$\varepsilon_\rho \leq \frac{d-r+1}{2}\ln(\rho+K) + \frac{(1-\rho+K)\kappa_{d+r-1}(1-\delta')}{2(\rho-K)}$$

where $K = \sqrt{\frac{1-\rho^2}{r}}t_r(1-\delta')$.

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 (**??**), 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 9), $p(y)$ 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 $\{Y : v^\top y > 0\}$. Analogously, $q(y)$ has the same form with v replaced by v', and is supported on $\{y : (v')^\top y > 0\}$. Define the support of q by

$$\mathcal{Z}_q := \{y : q(y) > 0\} = \{y : (v')^\top y > 0\}.$$

On $\mathcal{Z}_q$, define the privacy loss random variable for $y \sim q$,

$$L(y) = \ln \frac{p(y)}{q(y)}$$

Now for any measurable set $\mathcal{Y} \subset \mathbb{R}^d$ we have ,

$$p(\mathcal{Y}) = p(\mathcal{Y} \cap \mathcal{Z}_q) + p(\mathcal{Y} \cap \mathcal{Z}_q^c) \leq p(\mathcal{Z}_q^c) + \int_{\mathcal{Y} \cap \mathcal{Z}_q} p(y)dy \tag{14}$$

On $\mathcal{Z}_q$ we have $p(y) = e^{L(y)}q(y)$, so

$$\begin{aligned}
\int_{\mathcal{Y} \cap \mathcal{Z}_q} p(y)dy &= \int_{\mathcal{Y} \cap \mathcal{Z}_q} e^{L(y)} q(y)dy \\
&\overset{(a)}{\leq} e^{\varepsilon} q(\mathcal{Y} \cap \mathcal{Z}_q) + \int_{\mathcal{Y} \cap \mathcal{Z}_q} e^{L(y)} \mathbf{1}\{L(y) \geq \varepsilon\} q(y)dy \\
&\overset{(b)}{\leq} e^{\varepsilon} q(\mathcal{Y}) + \int_{\mathcal{Y} \cap \mathcal{Z}_q} e^{L(y)} \mathbf{1}\{L(y) \geq \varepsilon\} q(y)dy.
\end{aligned} \tag{15}$$

where step (a) follows by splitting the integrand into whether $L(y) < \varepsilon$ or $L(y) \geq \varepsilon$, and step (b) is due to $q(\mathcal{Y} \cap \mathcal{Z}_q) \leq q(\mathcal{Y})$.

Substituting Equation (15) into Equation (14), we get

$$\begin{aligned}
p(\mathcal{Y}) &\leq e^{\varepsilon} q(\mathcal{Y}) + p(\mathcal{Z}_q^c) + \int_{\mathcal{Z}_q} e^{L(y)} \mathbf{1}\{L(y) \geq \varepsilon\} q(y)dy \\
&= e^{\varepsilon} q(\mathcal{Y}) + p(\mathcal{Z}_q^c) + \int_{\mathcal{Z}_q} \mathbf{1}\{L(y) \geq \varepsilon\} p(y)dy \\
&\leq e^{\varepsilon} q(\mathcal{Y}) + p(\mathcal{Z}_q^c) + \mathbb{P}_{y \sim p}(L(Y) \geq \varepsilon, y \in \mathcal{Z}_q)
\end{aligned} \tag{16}$$

where the last inequality is due to $\mathcal{Y} \cap \mathcal{Z}_q \subset \mathcal{Z}_q$.

Next, we upper bound the second term $p(\mathcal{Z}_q^c)$. We have that

$$p(\mathcal{Z}_q^c) = \mathbb{P}_{y \sim p}((v')^\top y \leq 0) = \mathbb{P}_M((v')^\top M v \leq 0).$$

For $k \in [r]$ z_k be the kth column of Z, $z_k \sim \mathcal{N}(0, 1/rI_d)$. Let $v' = \rho v + \sqrt{1-\rho^2} w$ where $w \perp v, \|w\|_2 = 1$. Let the unit vector

$$w := \begin{cases} \dfrac{v' - \rho\, v}{\sqrt{1-\rho^2}}, & \text{if } |\rho| < 1, \\ \text{any unit vector in } v^{\perp}, & \text{if } |\rho| = 1. \end{cases}$$

Let $g_k := \sqrt{r} v^\top z_k$ and $h_k := \sqrt{r} w^\top z_k$. Then $g_k, h_k \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and are independent across k, and

$$(v')^\top Y = (v')^\top M v = \sum_{k=1}^{r} (v'^\top z_k)(v^\top z_k) = \frac{1}{r} \sum_{k=1}^{r} (\rho g_k + \sqrt{1-\rho^2}\, h_k) g_k.$$

Define

$$X := \sum_{k=1}^{r} g_k^2 \sim \chi_r^2, \qquad S := \sum_{k=1}^{r} g_k h_k.$$

Conditioned on $(g_1, \ldots, g_r)$, S is a linear combination of independent h_k's, so

$$S \mid (g_1, \ldots, g_r) \sim \mathcal{N}(0, X).$$

By rotational symmetry this implies $S \mid X = x \sim \mathcal{N}(0, x)$, Hence,

$$(v')^\top Y | (g_1, \ldots, g_r) \sim \mathcal{N}\left(\frac{\rho}{r}\sum_{k=1}^{r} g_k^2, \frac{1-\rho^2}{r^2}\mathrm{Var}(S|g_1, \ldots g_r)\right) \overset{d}{=} \mathcal{N}\left(\rho X, \frac{1-\rho^2}{r^2}X\right)$$

and similarly,

$$(v')^\top Y \mid X = x \sim \mathcal{N}\Big(\frac{\rho}{r}x,\ \frac{1-\rho^2}{r^2}x\Big),$$

and therefore

$$P\big((v')^\top Y \le 0 \mid X = x\big) = \Phi\Big(-\frac{\rho\sqrt{x}}{\sqrt{1-\rho^2}}\Big).$$

Taking expectation over $X \sim \chi_r^2$ yields

$$P\big((v')^\top Y \le 0\big) = \mathbb{E}_{X\sim\chi_r^2}\left[\Phi\Big(-\frac{\rho\sqrt{X}}{\sqrt{1-\rho^2}}\Big)\right]. \tag{17}$$

So we are left to choose an ε_ρ in a reasonable range so that $\mathbb{P}_{y\sim p}(L(y) \ge \varepsilon_\rho) \le \delta'$ lies in a reasonable range:

Assume $d > r$, for $y \in \mathcal{Z}_q$ the privacy loss random variable satisfies

$$\begin{aligned} L(y) = \log\frac{P(y)}{Q(y)} &= \frac{d-r+1}{2}\log\left(\frac{(v')^\top y}{v^\top y}\right) + \frac{r\,\|y\|^2}{2v^\top y}\left(\frac{v^\top y}{(v')^\top y} - 1\right) \\ &= \frac{d-r+1}{2}\log A + \frac{B}{2}\left(\frac{1}{A} - 1\right) \end{aligned} \tag{18}$$

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.By Lemma 8 we have

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

so for $\delta' > 0$, e, $t_r(\cdot)$ denote the quantile function of a student t distribution with degree of freedom r. Let

$$a_- = \rho - \sqrt{\frac{1-\rho^2}{r}}t_r\left(1 - \frac{\delta'}{3}\right), \quad a_+ = \rho + \sqrt{\frac{1-\rho^2}{r}}t_r\left(1 - \frac{\delta'}{3}\right)$$

Let $\kappa_{d+r-1}(\cdot)$ denote the quantile function of χ^2_{d+r-1}. Let $b = \kappa_{d+r-1}\left(1 - \frac{\delta'}{3}\right)$. Define the good set as $A \in (a_-, a_+)$, $B \le b$. As the good set implies an upper bound on $L(y)$,

$$\mathbb{P}_{y\sim p}\left[L(y) \le \frac{d-r+1}{2}\ln a_+ + \frac{b}{2}\left(\frac{1}{a_-} - 1\right)\right] \ge \mathbb{P}(\text{good set})$$

By the definition of quantile function,

$$\begin{aligned}\mathbb{P}(\text{good set}) &= \mathbb{P}(A \in (a_-, a_+) \cap B \le b) = 1 - \mathbb{P}(A \notin (a_-, a_+) \cup B > b) \\ &\ge 1 - \mathbb{P}(A \notin (a_-, a_+)) - \mathbb{P}(B > b) = 1 - \delta'.\end{aligned}$$

Thus, let $\varepsilon(a_-, a_+, b) \coloneqq \frac{d-r+1}{2} \ln \frac{1}{a_+} + \frac{b}{2}\left(\frac{1}{a_-} - 1\right)$,

$$\mathbb{P}_{y\sim p}\left[L(y) > \varepsilon(a_-, a_+, b)\right] \le 1 - \mathbb{P}(\text{good set}) \le \delta'$$

□

Lemma 8 (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 $y := Mv$. Define

$$A := \frac{(v')^\top y}{v^\top y}, \qquad B := \frac{r\|y\|^2}{v^\top y}.$$

Then there exist independent random variables

$$K_1 \sim \chi^2_r, \qquad K_2 \sim N(0,1), \qquad K_3 \sim \chi^2_{d-2},$$

such that, jointly,

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

In particular, if

$$T_r := \frac{K_2}{\sqrt{K_1/r}} \sim t_r$$

is Student-t with r degrees of freedom, then

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

Proof. Recall that $M = ZZ^\top/r$ where Z is d by r with $i.i.d$ Gaussian entries $N(0,1)$. 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 $K_2 \sim N(0,1)$, $K_1 \sim \chi_r^2$ 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 9 (PDF of Mv)**.** *Let $z_1, ..., z_r$ be i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$ where $d \geq r$, $M = \sum_{i=1}^r z_i z_i^T$, 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

$$a_i = z_i^\top v, \quad u_i = (I - vv^\top) z_i =: P_\perp z_i.$$

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

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

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

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

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

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

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

$$\mathbb{P}\left[S = s, Y_\perp = y_\perp\right] = \mathbb{P}\left[S = s\right]\mathbb{P}\left[Y_\perp = y_\perp|S = s\right] \tag{19}$$

$$f_{S\sigma^2}(s) = f_S\left(\frac{s}{\sigma^2}\right) = \frac{s^{r/2-1}e^{-\frac{s}{2\sigma^2}}}{2^{r/2}\Gamma(r/2)\sigma^{r-2}} \tag{20}$$

$$f_{Y_\perp|S}(y_\perp|S = s) = (2\pi)^{-\frac{d-1}{2}}\left(\sigma^2 s\right)^{-\frac{d-1}{2}}\exp\left(-\frac{\|y_\perp\|^2}{2\sigma^2 s}\right) \tag{21}$$

$$f_{S,Y_\perp}(s, y_\perp) = C_{r,d,\sigma}s^{\frac{r-d+1}{2}}e^{-\frac{s^2+\|y_\perp\|^2}{2\sigma^2 s}}, \quad C_{r,d,\sigma} = \frac{1}{2^{r/2}\Gamma(r/2)\sigma^{d-r-1}(2\pi)^{(d-1)/2}} \tag{22}$$

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

$$Y = Q^\top \begin{pmatrix} S \\ Y_\perp \end{pmatrix}. \tag{23}$$

By changing the variables from $(S, Y_\perp)$ to Y with Equation (23), we get the probablity density function for Y when $y^\top v \geq 0$ ($s > 0$), i.e.

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

$$\exp\left(-\frac{r}{2}\left(s + \frac{\left\|U^\top y\right\|^2}{v^\top y}\right)\right) = \exp\left(-\frac{r}{2v^\top y}\left(y^\top(vv^\top + UU^\top)y\right)\right) = \exp\left(-\frac{r\left\|y\right\|^2}{2v^\top y}\right) \tag{24}$$

So we get

$$f_Y(y) = C_{r,d,\sigma}(v^\top y)^{\frac{r-d+1}{2}}\exp\left(-\frac{\|y\|}{2\sigma^2 v^\top y}\right) \tag{25}$$

□

B.1 Vector Privacy Amplification and Applications

Privacy amplification by increasing effective alignment The privacy guarantees for random projection in **??** 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 10. *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).

B.2 Applications

In this section, we highlight three potential applications of the projection mechanism for the case. In Section 4, we highlight our main application, differentially privarte LoRA .

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, sample $M \sim \mathsf{W}_d\big(\sigma^2 I_d, r\big)$ once, and at each iteration update

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

This *Randomly Projected Gradient Descent (RP–GD)* algorithm retains directional information (which is what drives progress for many optimisers) while providing guaranteeing DP.

Private Retrival Another 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.

C Matrix Projection Mechanism is not private

In this section, we prove Lemma 11 that directly implies Proposition 2.

Lemma 11 (Almost-sure separation of images under random M)**.** *Let $V, V' \in \mathbb{R}^{d\times m}$ with $\Delta V := V - V' \neq 0$, and let $M = ZZ^\top$ where $Z \in \mathbb{R}^{d\times r}$ has i.i.d. $\mathcal{N}(0,1)$ entries. Then*

$$\mathbb{P}\big(MV = MV'\big) \;=\; \mathbb{P}\big(M\Delta V = 0\big) \;=\; 0.$$

In particular, the two random images $\{MV : M\}$ and $\{MV' : M\}$ intersect only on a $\mathbb{P}$-null set (with randomness over M).

Proof. We have $MV = MV'$ iff $M\Delta V = 0$. Since $M = ZZ^\top$ is positive semidefinite, for any vector x,

$$ZZ^\top x = 0 \quad\Longleftrightarrow\quad x^\top ZZ^\top x = \|Z^\top x\|_2^2 = 0 \quad\Longleftrightarrow\quad Z^\top x = 0.$$

Applying this columnwise shows $M\Delta V = 0 \iff Z^\top \Delta V = 0$.

Let $s = \operatorname{rank}(\Delta V) \geq 1$ and write $\Delta V = UB$ where $U \in \mathbb{R}^{d\times s}$ has orthonormal columns and $B \in \mathbb{R}^{s\times m}$ has full row rank. Then

$$Z^\top \Delta V = 0 \implies (Z^\top U)B = 0 \implies Z^\top U = 0,$$

since B has full row rank. But $Z^\top U \in \mathbb{R}^{r\times s}$ has i.i.d. $\mathcal{N}(0,1)$ entries (orthogonal invariance), hence $\mathbb{P}(Z^\top U = 0) = 0$. Therefore $\mathbb{P}(M\Delta V = 0) = 0$. □

C.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:

$$W_{t+1} - W_t = -\eta \nabla_W L(W_t)(A^\top A) \tag{26}$$

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:

$$W_1 - W_0 = -\eta \nabla_W L(W_0) A_0^\top A_0 \tag{27}$$

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 Section 3.2, which establishes that such projections are not private. Therefore, we conclude that LoRA is not private without additive noise.

D Privacy Analysis of Matrix Projection Mechanism

D.1 Proofs 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

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

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

Lemma 12 (Exact orthogonal split)**.** *We have $M = M_\parallel + M_\perp$ and $Z_\parallel Z_\perp^\top = 0$. 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, $Z_\parallel = ZP_H$ and $Z_\perp = ZP_H^\perp$. Therefore,

$$M \ = \ ZZ^\top \ = \ Z(P_H + P_H^\perp)(P_H + P_H^\perp)^\top Z^\top.$$

Using symmetry of the projectors and expanding, we obtain

$$\begin{aligned} M &= Z(P_H + P_H^\perp)(P_H + P_H^\perp)Z^\top \\ &= ZP_HZ^\top + ZP_H^\perp Z^\top + ZP_HP_H^\perp Z^\top + ZP_H^\perp P_H Z^\top \\ &= ZP_HZ^\top + ZP_H^\perp Z^\top \ = \ Z_\parallel Z_\parallel^\top + Z_\perp Z_\perp^\top \ = \ M_\parallel + M_\perp, \end{aligned}$$

since $P_HP_H^\perp = P_H^\perp P_H = 0$. This proves $M = M_\parallel + M_\perp$.

Next, the cross term vanishes:

$$Z_\parallel Z_\perp^\top \ = \ (ZP_H)(ZP_H^\perp)^\top \ = \ ZP_H(P_H^\perp)^\top Z^\top \ = \ ZP_HP_H^\perp Z^\top \ = \ 0.$$

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

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

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 $GP_H^\perp = 0$, 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

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

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 13 (Posterior stability of the residual block)**.** *Conditional on G (and hence on H and P_H), the random matrix $Z_\perp = ZP_H^\perp$ is independent of Y. Equivalently,*

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

Proof. Recall the orthogonal decomposition

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

where G and W are independent and have i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries. Let $H = \mathrm{rowspan}(G)$ and let $P_H^\perp$ be the orthogonal projector onto $H^\perp$. By definition,

$$Z_\perp = ZP_H^\perp = (UG + U_\perp W)P_H^\perp = U(GP_H^\perp) + U_\perp(WP_H^\perp).$$

Since $H = \mathrm{rowspan}(G)$, every row of G lies in H, hence projecting onto $H^\perp$ annihilates the rows of G, i.e. $GP_H^\perp = 0$. Therefore

$$Z_\perp = U_\perp \, W \, P_H^\perp. \tag{28}$$

In particular, conditional on G (and therefore conditional on H and P_H), the projector $P_H^\perp$ is deterministic, and equation 28 shows that $Z_\perp$ is a measurable function of W only.

Next, by the Gaussian block decomposition above, the matrices

$$G \coloneqq U^\top Z \in \mathbb{R}^{s \times r} \qquad \text{and} \qquad W \coloneqq U_\perp^\top Z \in \mathbb{R}^{(d-s) \times r}$$

are independent and have i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries.

Recall that $M = ZZ^\top$ and that we defined

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

Since $v_k \in S = \mathrm{span}(V_{-j})$ for all $k \neq j$, Lemma 12 gives $M_\perp v_k = 0$, and therefore

$$Mv_k = (M_\parallel + M_\perp)v_k = M_\parallel v_k \qquad \text{for all } k \neq j.$$

Hence Y can be written as

$$Y = \left[\, M_\parallel v_k + \xi_k \,\right]_{k \neq j} = \left[\, Z_\parallel Z_\parallel^\top v_k + \xi_k \,\right]_{k \neq j},$$

which shows that Y depends on Z only through $Z_\parallel = ZP_H$ and the independent noises $\{\xi_k\}_{k \neq j}$. Moreover, conditional on G (hence on H and P_H), we can make the dependence on W explicit. Since

$$Z_\parallel = ZP_H = (UG + U_\perp W)P_H = UG + U_\perp(WP_H),$$

it follows that, conditional on G, the random variable Y is measurable with respect to $\sigma(WP_H, \{\xi_k\}_{k \neq j})$. On the other hand,

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

so $Z_\perp$ is measurable with respect to $\sigma(WP_H^\perp)$ (conditional on G). Finally, conditional on G, the Gaussian matrix W decomposes as

$$W = WP_H + WP_H^\perp,$$

where WP_H and $WP_H^\perp$ are independent. Therefore,

$$Z_\perp \;\perp\!\!\!\perp\; Y \mid G.$$

Finally, since $M_\perp = Z_\perp Z_\perp^\top$ is a measurable function of $Z_\perp$, the same conditional independence carries over:

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

which concludes the proof.

□

Lemma 2 (Residual DP bound)**.** *Write $s \coloneqq \dim(S)$ and $m \coloneqq d - s$. Fix $\rho_\perp \in (0, 1]$ and $\delta' \in (0, 1)$, and define*

$$K_\perp = \sqrt{\tfrac{1-\rho_\perp^2}{q}}\, t_q\left(1 - \delta'/3\right),$$

Then the residual release $R(b) = M_\perp b$ is $(\varepsilon_\perp, \delta_\perp)$-DP for inputs b, b' satisfying Definition 4 where

$$\delta_\perp = \mathbb{E}_{X \sim \chi^2_q}\left[\Phi\left(-\tfrac{\rho_\perp \sqrt{X}}{\sqrt{1-\rho_\perp^2}}\right)\right] + \delta',$$

$$\varepsilon_\perp \leq \tfrac{m-q+1}{2} \ln(\rho_\perp + K_\perp) + \frac{\kappa_{m+q-1}\left(1 - \delta'/3\right)}{2}\left(\tfrac{1}{\rho_\perp - K_\perp} - 1\right).$$

Proof. We treat the two cases separately.

Case 1: $\beta_\perp = 0$**.** We defined $\|b\| = \beta_\perp$ therefore $\beta_\perp = 0$ means b=0

$$R(b) = M_\perp b = 0 \qquad \text{and similarly} \qquad R(b') = M_\perp b' = 0$$

deterministically.

Case 2: $\beta_\perp > 0$**.** Let $G = U^\top Z$ and recall that $H = \text{rowspan}(G)$ and $Z_\perp = ZP_H^\perp$. By Lemma 13

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

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

Fix G (equivalently, 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 $b \in \mathbb{R}^d$,

$$R(b) = M_\perp b \;=\; U_\perp\, (WP_H^\perp)(WP_H^\perp)^\top \underbrace{(U_\perp^\top b)}_{=:\, b_\perp}. \tag{29}$$

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 $H = \text{rowspan}(G)$ 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 $p = \dim(H) = \text{rank}(G)$.

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 29 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 14. *Fix $\beta \in (0,1)$ and set $g_\beta = \sqrt{2\ln(2/\beta)}$. Conditional on H, with probability at least $1-\beta$ over the draw of Z,*

$$\big\|M_\parallel u\big\| \le \sigma_M^2 \left(\sqrt{d} + \sqrt{p} + g_\beta\right)(\sqrt{p} + g_\beta) \coloneqq \Gamma_\beta. \tag{30}$$

Proof. Let P_H be the orthogonal projector onto H and recall that

$$Z_\parallel = ZP_H, \qquad M_\parallel = Z_\parallel Z_\parallel^\top.$$

Fix a unit vector $u \in \mathbb{R}^d$ (the bound scales by $\|u\|$ otherwise). Conditional on H, choose an orthonormal basis matrix $Q \in \mathbb{R}^{r\times p}$ for H so that

$$P_H = QQ^\top, \qquad Q^\top Q = I_p.$$

Define the $d \times p$ Gaussian matrix

$$\widetilde{Z} \;\coloneqq\; ZQ.$$

Then

$$M_\parallel = ZP_H Z^\top = ZQQ^\top Z^\top = (ZQ)(ZQ)^\top = \widetilde{Z}\,\widetilde{Z}^\top,$$

and hence

$$\|M_\parallel u\| = \|\widetilde{Z}\,\widetilde{Z}^\top u\| \le \|\widetilde{Z}\|_{\mathrm{op}} \cdot \|\widetilde{Z}^\top u\|. \tag{31}$$

Since each row of Z is distributed as $\mathcal{N}(0, \sigma_M^2 I_r)$ and Q has orthonormal columns, we have for each row $z_i^\top$ of Z,

$$(z_i^\top Q)^\top \sim \mathcal{N}(0, \sigma_M^2 I_p).$$

Rows remain independent, hence conditional on H, $\widetilde{Z} \in \mathbb{R}^{d\times p}$ has i.i.d. $\mathcal{N}(0, \sigma_M^2)$ entries.

Let $G \in \mathbb{R}^{d\times p}$ have i.i.d. $\mathcal{N}(0,1)$ entries so that $\widetilde{Z} = \sigma_M G$. A standard Gaussian operator norm bound gives that for all $t \ge 0$,

$$\mathbb{P}\Big(\|G\|_{\mathrm{op}} \ge \sqrt{d} + \sqrt{p} + t\Big) \le e^{-t^2/2}.$$

Moreover, since u is fixed and $\|u\| = 1$, we have $G^\top u \sim \mathcal{N}(0, I_p)$, and thus

$$\mathbb{P}\Big(\|G^\top u\| \ge \sqrt{p} + t\Big) \le e^{-t^2/2}.$$

Set $t = g_\beta = \sqrt{2\ln(2/\beta)}$. Then $e^{-t^2/2} = \beta/2$, and scaling back by σ_M yields

$$\mathbb{P}\Big(\|\widetilde{Z}\|_{\mathrm{op}} \le \sigma_M(\sqrt{d} + \sqrt{p} + g_\beta) \,\Big|\, H\Big) \ge 1 - \beta/2,$$
$$\mathbb{P}\Big(\|\widetilde{Z}^\top u\| \le \sigma_M(\sqrt{p} + g_\beta) \,\Big|\, H\Big) \ge 1 - \beta/2.$$

By a union bound, with conditional probability at least $1 - \beta$ (given H), both events hold.

All together this mean with probability at least $1 - \beta$ (given H) we have

$$\|M_\parallel u\| \le \|\widetilde{Z}\|_{\mathrm{op}} \cdot \|\widetilde{Z}^\top u\| \le \sigma_M^2(\sqrt{d} + \sqrt{p} + g_\beta)(\sqrt{p} + g_\beta) = \Gamma_\beta.$$

□

Lemma 15. *Fix $\beta \in (0,1)$ 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 $(\varepsilon_\parallel, \delta_\parallel + \beta)$-DP, where*

$$\varepsilon_\parallel \;=\; \frac{\Gamma_\beta \Delta_v}{\sigma_G}\sqrt{2\ln\Big(\frac{1.25}{\delta_\parallel}\Big)},$$

where Γ_β is as in Lemma 14.

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

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

By Lemma 14, we have $\Pr\left(\mathcal{E}_{u,\beta}\right) \ge 1 - \beta$.

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

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

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

$$\left\| M_\parallel \left\| v - v' \right\| u \right\| \le \Gamma_\beta \Delta_v$$

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

$$\varepsilon_\parallel = \frac{\Gamma_\beta \Delta_v}{\sigma_G}\sqrt{2\ln\left(\frac{1.25}{\delta_\parallel}\right)}.$$

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

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

□

Theorem 4. *Fix $j \in [n]$. Let $V \sim_A^{(j)} V'$ 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 (3). 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*

$$\varepsilon_{\text{one}} \le \varepsilon_\parallel + \varepsilon_\perp, \qquad \delta_{\text{one}} \le \delta_\parallel + \delta_\perp + \delta_s.$$

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

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

So if we define

$$R(b) = M_\perp b$$
$$C(v) = M_\parallel v_j$$

we know by Lemma 12, 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 15 establishes $(\varepsilon_\parallel, \delta_\parallel)$-differential privacy using the Gaussian mechanism with the directional sensitivity bound from Lemma 14.

By composition and the postprocessing Lemma this tells us that conditioned on H and $\|M_\parallel u\| \le \Gamma_\beta$ (X, Y) is $(\varepsilon_\parallel + \varepsilon_\perp, \delta_\parallel + \delta_\perp)$-DP. We can remove the conditioning on $\|M_\parallel u\| \le \Gamma_\beta$ by adding an additional β 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. □

Proof. Write the mechanism as $\mathcal{A}(V) = MV + \Xi$, and denote its jth column by $X \coloneqq Mv_j + \xi_j$ and the remaining columns by $Y \coloneqq \left[\, Mv_k + \xi_k \,\right]_{k \neq j}$.

Using the orthogonal split $M = M_\parallel + M_\perp$ and Lemma 12, we have $M_\perp v_k = 0$ for all $k \neq j$. Hence

$$Y = \left[\, M_\parallel v_k + \xi_k \,\right]_{k \neq j}, \qquad X = M_\parallel v_j + \xi_j + M_\perp v_j.$$

Define the residual map $R(b) \coloneqq M_\perp b$ and the "main" (correlated) map

$$\mathcal{C}(v) \coloneqq M_\parallel v + \xi_j.$$

We will first argue about the privacy of X conditioned on Y and in a final step remove the conditioning.

By Lemma 13, conditional on H the residual block $M_\perp$ is independent of Y. So conditioned on (H, Y) $M_\perp$ is distributed like a random Wishart random matrix which is what Lemma 2 exploits to show that conditional on (H, Y), the map $b \mapsto R(b) = M_\perp b$ is $(\varepsilon_\perp, \delta_\perp)$-DP.

By Lemma 15, conditional on (H, Y), the mechanism

$$v \mapsto \mathcal{C}(v) = M_\parallel v + \xi_j$$

is $(\varepsilon_\parallel, \delta_\parallel + \beta)$-DP with

$$\varepsilon_\parallel = \frac{\Gamma_\beta \Delta_v}{\sigma_G} \sqrt{2 \ln\Big(\frac{1.25}{\delta_\parallel}\Big)}.$$

Conditional on (H, Y) and on E_β, the release of X can be written as the composition of two DP mechanisms

$$X = \mathcal{C}(v_j) + R(v_j),$$

is therefore by sequential composition $(\varepsilon_\parallel + \varepsilon_\perp, \delta_\parallel + \delta_\perp)$-DP conditional on (H, Y, E_β).

By Lemma 14, $\mathbb{P}(\mathsf{E}_\beta^c \mid H) \leq \beta$. Thus the same mechanism is $(\varepsilon_\parallel + \varepsilon_\perp, \delta_\parallel + \delta_\perp + \beta)$-DP conditional on (H, Y).

Since $V \sim_A^{(j)} V'$ implies $V_{-j} = V'_{-j}$, the random variables H and Y are distributed the same under V and V'. Therefore, a conditional DP guarantee for X given (H, Y) implies that the joint release (X, Y) is (ε, δ)-DP with

$$\varepsilon \leq \varepsilon_\parallel + \varepsilon_\perp, \qquad \delta \leq \delta_\parallel + \delta_\perp + \beta.$$

□

D.2 Proofs 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 16. *For $L_M(Y)$ defined as*

$$L_M(Y) := \frac{dP(\cdot|M)}{dQ(\cdot|M)}(Y)$$

we have

$$\log L_M(Y)|M \sim \mathcal{N}\left(-\frac{\|P_M \Delta V\|_F^2}{2\sigma^2}, \frac{\|P_M \Delta V\|_F^2}{\sigma^2}\right)$$

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:

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

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:

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

Now let's recall that

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

so the jth column is distributed as

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

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

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

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

$$\begin{aligned}
(\alpha A)^\dagger &= \frac{1}{\alpha}A^\dagger \\
(A\otimes B)^\dagger &= A^\dagger \otimes B^\dagger
\end{aligned}$$

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 finally get

$$(\Delta\mu)^\top \Sigma^\dagger (\Delta\mu) = \|P_M \Delta V\|_F^2.$$

Plugging this into the mean/variance formulas we get

$$\log L_M(Y)|M \sim \mathcal{N}(-\frac{\|P_M\Delta V\|_F^2}{2\sigma^2}, \frac{\|P_M\Delta V\|_F^2}{2\sigma^2})$$

□

Lemma 17 (Tail bound for Haar projection)**.** *Let $r \in \{1, \dots, d-1\}$ 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 17 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,

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

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 $\varepsilon > 0$, the mechanism $\mathcal{A}_2$ is $(\varepsilon, \delta_\alpha(\varepsilon))$-DP with*

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

where $I_\alpha(a,b)$ is the regularized incomplete Beta function (i.e. $I_\alpha(a,b) = \Pr[B \leq \alpha]$ for $B \sim \mathrm{Beta}(a,b)$).

Proof. First note that

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

So 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

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

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

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

Therefore,

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

Combine with the δ_M bound for the complement to obtain

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

Next Lemma 17 gives us a bound on δ_M and finally Fix $\varepsilon > 0$. For each fixed M, define the conditional "good output set"

$$\mathcal{Y}_\varepsilon(M) \;:=\; \big\{\, y : |\log L_M(y)| \leq \varepsilon \,\big\}.$$

On $\mathcal{Y}_\varepsilon(M)$ we have the pointwise bound

$$e^{-\varepsilon} \;\leq\; L_M(y) \;\leq\; e^{\varepsilon}.$$

Moreover, since $\ell_M(Y)$ is Gaussian as above, we can write its two-sided tail exactly in terms of the standard normal CDF Φ:

$$\mathbb{P}(Y \notin \mathcal{Y}_\varepsilon(M) \mid M) \;=\; \mathbb{P}(|\ell_M(Y)| > \varepsilon \mid M) \;=\; \Phi\left(\frac{-\varepsilon - \mu(M)/2}{\sqrt{\mu(M)}}\right) \;+\; 1 - \Phi\left(\frac{\varepsilon - \mu(M)/2}{\sqrt{\mu(M)}}\right),$$

with the convention that if $\mu(M) = 0$ then this probability equals 0 (indeed $\ell_M(Y) = 0$ almost surely).

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

$$G_\alpha \;:=\; \big\{\, M : \|P_M \Delta V\|_F^2 \leq \alpha\,\|\Delta V\|_F^2 \,\big\}.$$

On G_α, we have the uniform bound

$$\mu(M) \;\leq\; \bar{\mu}, \qquad \text{where} \qquad \bar{\mu} \;:=\; \frac{\alpha\,\|\Delta V\|_F^2}{\sigma^2}.$$

Since the tail expression above is increasing in $\mu(M)$ for the relevant regime, we can upper bound it by the same expression with $\mu(M)$ replaced by $\bar{\mu}$. Define

$$\delta_E(\varepsilon, \alpha) \;:=\; \Phi\!\left(\frac{-\varepsilon - \bar{\mu}/2}{\sqrt{\bar{\mu}}}\right) \;+\; 1 - \Phi\!\left(\frac{\varepsilon - \bar{\mu}/2}{\sqrt{\bar{\mu}}}\right), \qquad \bar{\mu} = \frac{\alpha \,\|\Delta V\|_F^2}{\sigma^2}.$$

Then, for all $M \in G_\alpha$,

$$\mathbb{P}(Y \notin \mathcal{Y}_\varepsilon(M) \mid M) \;\le\; \delta_E(\varepsilon, \alpha).$$

□

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.,

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

Remark 2. *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

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

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

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

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

$$T(\varepsilon; \alpha_0\mu) \;=\; \frac{1}{2}\, T(\varepsilon; \mu) \;=\; \frac{1}{2}\, \delta_{\text{Gauss}}(\varepsilon). \tag{32}$$

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

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

Assume additionally that $r \le d/2$, so that $\alpha \in (0, 1)$ and Lemma 7 applies. Then

$$s\left[1 - I_\alpha\!\left(\frac{r}{2}, \frac{d-r}{2}\right)\right] \;=\; s\,\Pr(B > \alpha) \;\le\; 2s \exp\!\left(-\frac{\eta^2 r}{72}\right), \tag{33}$$

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

We impose two conditions on r:

$$2s \exp\left(-\frac{\eta^2 r}{72}\right) \le \frac{1}{2}\,\delta_{\mathrm{Gauss}}(\varepsilon), \tag{34}$$

$$\alpha = (1+\eta)\frac{r}{d} \le \alpha_0. \tag{35}$$

Condition equation 35 is equivalent to $r \le \frac{\alpha_0}{1+\eta}\, d$, which is ensured whenever $r \ll d$. Condition equation 34 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 32,

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

and by equation 33 and equation 34,

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

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. □

E LoRA

Algorithm 1 One LoRA step with frozen A (single layer)

Input: pretrained weights $W_0 \in \mathbb{R}^{n\times d}$; rank $r < \min\{n, d\}$; dataset size N; loss $\mathcal{L}$; step size η; minibatch size B_{mb}.

Sample minibatch $\mathcal{B} \subset [N]$ with $|\mathcal{B}| = B_{\mathrm{mb}}$ (e.g. Poisson rate $q = B_{\mathrm{mb}}/N$).
Sample $A \sim \mathcal{N}(0, 1/r)^{r\times d}$ and freeze it.
Initialize $B \leftarrow 0 \in \mathbb{R}^{n\times r}$.
Form effective weights for this step: $W \leftarrow W_0 + BA$.
Compute gradient $G \leftarrow \nabla_B \mathcal{L}\big(W, \mathcal{B}\big) \in \mathbb{R}^{n\times r}$.
Update $B \leftarrow B - \eta\, G$.
Subsequent forward passes use $W = W_0 + BA$.

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 $B_t = B_{t-1} - \eta\, \nabla_B \mathcal{L}(W; \mathcal{B}_t)$.

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

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

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 2 describes one step of one layer of DP-LoRA-FA.

Algorithm 2 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 $|\mathcal{B}_t| = B_{\text{mb}}$ (e.g. Poisson rate $q = B_{\text{mb}}/N$).
  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: $\widehat{G}_t \leftarrow \frac{1}{B_{\text{mb}}} \sum_{i\in\mathcal{B}_t} \widetilde{G}_{t,i} \;+\; \frac{\sigma}{B_{\text{mb}}} E_t$, 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 $W = W_0 + BA$).

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:

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

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. For mutlistep and multilayer implementation details please se Section F.

Algorithm 3 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'; privacy parameters (ε, δ).

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

Set $B \leftarrow 0 \in \mathbb{R}^{n \times r}$

Sample $A \sim \mathcal{N}(0, 1/r)^{r \times d}$

Set $W_{\text{eff}} \leftarrow W_0 + BA$

Sample minibatch $\mathcal{B}_t \subset [N]$ with $|\mathcal{B}_t| = B_{\text{mb}}$ (or take $\mathcal{B}_t = [N]$).

Compute gradient $G \leftarrow \nabla_W \mathcal{L}(W_{\text{eff}}; \mathcal{B}_t) \in \mathbb{R}^{n \times d}$.

DP gradient: $\widehat{G} \leftarrow (\text{clip}_{\beta'}(G) + \sigma' E') A^\top A$, where $E'_{ij} \sim \mathcal{N}(0, 1)$ i.i.d.

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

In order to compare Algorithm 3 to Algorithm 2 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:

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

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 18 (Johnson-Lindenstrauss [Dasgupta and 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

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

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

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

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

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

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

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

equivalently (by row-stacking),

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

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

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

were we recall $r < d$. We can use Lemma 6 to bound the eigenvalues of C and hence the amount of noise we add. Note that Lemma 6 is stated for $W = \frac{1}{r} Z Z^\top$, 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 $\frac{1}{d} Z Z^\top$. Since $\frac{1}{d} Z Z^\top = \frac{r}{d} W$, the eigenvalues rescale by the factor r/d, turning the $\sqrt{d/r}$ scale in Lemma 6 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 $C = \frac{r}{d} A A^\top$ 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

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

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 $d \gg \log(1/\gamma)$).

E.2 DP Projection Mechanism without clipping

If we are in a setting where we know we do not have to clip, as the gradients are naturally bounded then we can rewrite the projection mechanism for LoRA so that only the gradients with respect to B are being used. Allowing us to regain computational and memory efficiency, which is one of the main motivations of LoRA.

Algorithm 4 Projection Mechanism without clipping

Input: Pre-trained model parameters $W_0 \in \mathbb{R}^{n\times d}$, low dimension r with $r < d, n$, training data $x \in \mathbb{R}^d$, loss function $\mathcal{L}$, number of rounds T, step size η, privacy noise $\sigma > 0$

Initialize $B_0 \in \mathbb{R}^{n\times r}$ and $A \in \mathbb{R}^{r\times d}$ randomly
for $t = 1, \dots T-1$ **do**
  $W_t \leftarrow W_0 + B_{t-1}A$ {update model}
  $y_t \leftarrow W_t x$ {evaluate}
  $G_t \leftarrow \frac{\partial \mathcal{L}(y)}{\partial B}|_{y=y_t}$ {calculate gradient}
  $B_t = B_{t-1} + \eta g_t$ {update B}
end for
$E \sim \mathcal{MN}(0, I_d, I_k)$ {sample noise}
$\tilde{G} \leftarrow \sum_{t=1}^{T} G_t + \sigma E A^\top$ {privatize gradients}
$\tilde{W} \leftarrow W_0 - \eta \tilde{G} A$

F Experiment details

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

F.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:

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

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 $\frac{1}{2}(\text{TPR} + \text{TNR})$.

(a) **AUC**

Noise	**16**	**64**	**128**	**256**	**384**	**512**
0.1	0.78	0.91	0.96	0.99	1.00	1.00
0.5	0.56	0.70	0.74	0.76	0.82	0.86

(b) **Balanced accuracy**

Noise	**16**	**64**	**128**	**256**	**384**	**512**
0.1	0.74	0.83	0.92	0.98	0.98	0.98
0.5	0.56	0.67	0.70	0.71	0.75	0.80

(c) **Test accuracy (%)**

Noise	**16**	**64**	**128**	**256**	**384**	**512**
0.1	44.61	49.98	54.52	57.94	60.38	61.70
0.5	33.13	37.14	42.57	46.48	48.26	50.92

Table 2: Performance vs. projection rank r for two noise levels.

	Rank r					
Metric	**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 (200 trials) vs. projection rank r at noise multiplier 0.5 ($\varepsilon \approx 136.05$, $\delta = 10^{-5}$).

Additional results on Noisy projection mechanism We then run noisy LoRA via gradient descent with $M(V + E)$, matching the small-r regime in Section 4.2. Results based on training 200 INOUT 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,

F.2 Experimental details for the comparison between equation M2, DP-LoRA, and DP-SGD

Setting. We follow the representation-learning setup of Pinto et al. [2024] and use their fixed feature extractor: 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 $y_i \in \{0, \ldots, 9\}$. 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 2), and (iii) our noisy-projection mechanism equation M2 (Algorithm 3) 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.

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, $\bigcap_{t=1}^{T}\mathcal{G}(M_t)$, 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 largest 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 all three methods we use batch size 1024. 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.

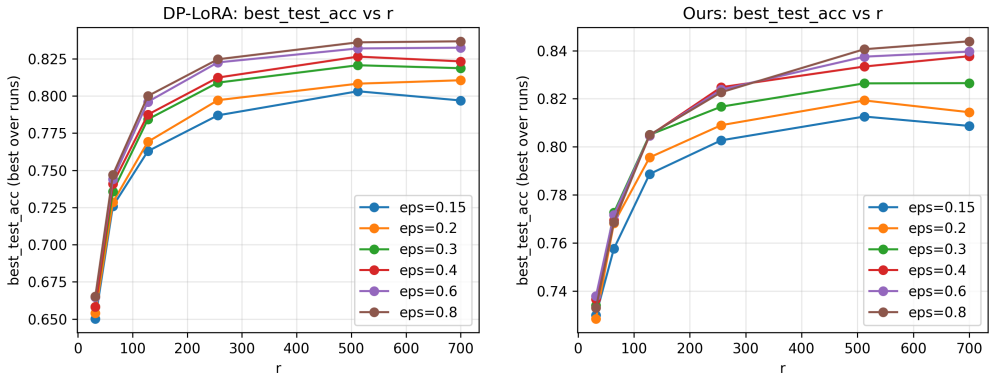

Figure 4: Best (non-private) test accuracy as a function of the rank r for DP-LoRA-FA (left) and our noisy-projection mechanism (right), under the same target privacy budget used in the main comparison.