Beyond Additive Noise: DP for LoRA via Random Projections

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

[1]Department of Empirical Inference, Max Planck Institute
{yaxi.hu, bernhard.schoelkop}@tuebingen.mpg.de
[2]Department of Computer Science, University of Copenhagen
{jodu, amsa}@di.ku.dk

Abstract

We study the differential privacy (DP) of low-rank adaptation (LoRA) fine-tuning. Focusing on FA-LoRA (fixed A, trained B), where a single training step is equivalent to applying a random Wishart projection to the gradients, we prove a formal (ε, δ)-DP guarantee for LoRA without adding explicit additive noise. The resulting privacy parameters depend explicitly on dataset sensitivity and the projection rank r. Moreover, the low-rank structure reduces memory and computation by design. To place these results in a broader context, we formalize the underlying projection operation as a general *projection mechanism* of which LoRA is an instance. This mechanism is of independent interest as random projections are ubiquitous in machine learning.

1 Introduction

Differential Privacy (DP) is widely regarded as the gold standard for protecting training data 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, the private counterpart of the workhorse 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 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., 2023], quantization-aware tuning for low-bit backbones [Dettmers et al., 2023, Li et al., 2023], stability/initialization refinements [Hayou et al., 2024, Meng et al., 2024], and structural decompositions [Liu et al., 2024], each targeting stronger quality under tight compute/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 existing privatisation algorithms largely ignore this inherent randomness in their algorithmic design. At the same time, empirical studies report reduced memorisation under LoRA [Hong et al., 2025] (without any explicit privatisation) and note training dynamics that close match that of DP-SGD [Malekmohammadi and Farnadi, 2025]. These observations suggest that the built-in randomness may play a central role in privacy, raising the possibility that LoRA could be *provably private by design*. To our knowledge, no prior work establishes a formal DP guarantee for LoRA.

A formal DP guarantee for LoRA *without* additive noise offers a practical route through the privacy–accuracy–compute trilemma: it can lower computational overhead while preserving accuracy and ensuring privacy. In this work, we show that LoRA is provably differentially private.

Our key observation (see Section 2) is that certain LoRA variants (e.g. FA-LoRA [Hao et al., 2024, Sun et al., 2024] with A fixed and only B trained) update parameters as if they applied a random Wishart projection to the gradient. Leveraging this equivalence, we prove that FA-LoRA is (ε, δ)-DP. To place this in a broader context, we formalize the underlying operation as a general *projection mechanism*, which multiplies the output by a Wishart-distributed random matrix and establish its privacy guarantees.

Definition 1 (Projection mechanism)**.** *Let $\mathcal{S} \subset \mathcal{X}^n$ be a dataset collection and $f : \mathcal{S} \to \mathbb{R}^{d\times m}$ a query function. For $r \in \mathbb{N}$ and $\sigma^2 > 0$, the (Wishart) projection mechanism is defined by*

$$\mathcal{A}_{r,\sigma^2}(S) \;=\; M\, f(S), \qquad M := ZZ^\top,$$

where $Z \in \mathbb{R}^{d\times r}$ has independent columns $z_k \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2 I_d)$ or equivalently, $M \sim \mathsf{W}_d(\sigma^2 I_d,\, r)$.

We believe this projection-based approach is of independent interest: it departs from the classical additive-noise mechanisms(e.g. Lemma 1) yet yields rigorous privacy guarantees. Random projections, such as Wishart transforms, are already common in standard ML (e.g., dimensionality reduction, sketching, randomized preconditioning), and we expect that this mechanism will have uses beyond LoRA. At this stage we also note that this is different from the common JL transformation: the expectation of the JL transformation matrix is zero, whereas the expectation of the Wishart transformation matrix is a low rank identity matrix. Intuitively, JL preserves the norm of a vector whereas Wishart preserves the angle.

Our Contributions. Our main contributions are twofold. First, we show that LoRA (and FA-LoRA) is differentially private, with privacy parameters controlled by properties of the data and architecture, concretely, by the ratio of the largest to smallest singular values of the gradient matrix. Second, to obtain this result, we introduce and analyse a new DP mechanism: the random Wishart projection mechanism, and establish its DP guarantees, discuss simple privacy amplification techniques, and outline applications beyond LoRA.

Organization. In Section 2 we review DP preliminaries and give a brief overview of why LoRA is private. Section 3 establishes privacy and convergence guarantees for the projection mechanism with vector outputs ($m = 1$) and presents a simple privacy amplification method. Section 4 extends these results to matrix-valued outputs ($m > 1$) and derives a privacy guarantee for LoRA. Finally, Section 5 covers related work, limitations, and open questions.

2 Preliminaries and Main Implications for LoRA

Before presenting our main results, we recall basic differential privacy (DP) notions and composition tools we rely on. We then introduce LoRA, explain its equivalence to the projection mechanism (Definition 1), and state the per-step privacy guarantee enjoyed by LoRA.

Differential privacy limits how much the output distribution can change when a single data point is modified.

Definition 2 (Neighboring datasets)**.** *Let S, S' be datasets of same size. We write $S \sim_H S'$ if they differ in exactly one entry, i.e. $d_H(S, S') = 1$, where d_H is the Hamming distance.*

Then, differential privacy can be defined as follows.

Definition 3 (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: i.e. by adding noise to the output of a non-private query f, calibrated to the sensitivity of the underlying query f.

Definition 4 (Sensitivity)**.** *For $f : \mathcal{S} \to \mathbb{R}^d$, its ℓ_2 sensitivity is*

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

One such example of canonical additive mechanisms is the Gaussian mechanism

Lemma 1 (Gaussian mechanism)**.** *Let $f : \mathcal{S} \to \mathbb{R}^d$ have sensitivity Δ. The mechanism*

$$\mathcal{A}(S) = f(S) + Z, \qquad Z \sim \mathcal{N}\left(0, \frac{2\Delta^2 \log \frac{2}{\delta}}{\varepsilon^2} I_d\right)$$

is (ε, δ)-DP.

Restricting the domain of $\mathcal{A}$ to a dataset collection $\mathcal{D} \subseteq \mathcal{S}$ yields a DP guarantee *conditioned on* $\mathcal{D}$. In practice, this often reduces Δ and improving utility but comes at the cost of not providing privacy on datasets outside the collection. Several approaches have been proposed in the literature to check whether a dataset indeed belongs to a collection $\mathcal{D}$, including the Propose-Test-Release mechanism (PTR) Dwork and Lei [2009].

Basic Properties of DP DP mechanisms admit several simple but useful properties, including composition and amplification via subsampling, as stated below.

Lemma 2 (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 (Poisson subsampling)**.** *Given a dataset $D = \{x_1, \ldots, x_N\}$, include each x_i independently with probability $q \in (0, 1)$ to form a subsample S. Equivalently, draw $m_i \sim \text{Bernoulli}(q)$ i.i.d. and set $S = \{x_i : m_i = 1\}$. Then $|S| \sim \text{Binomial}(N, q)$ and $\mathbb{E}[|S|] = qN$.*

Lemma 3 (Amplification by subsampling)**.** *Let $\mathcal{A}$ be (ε, δ)-DP. Under Poisson subsampling with rate q, the composed mechanism $\mathcal{A} \circ S_q$ is $\big(\log(1 + q(e^{\varepsilon} - 1)), q\delta\big)$-DP.*

Algorithm 1 One LoRA step with frozen A

Input: number of layers $\ell_{\max}$, pretrained weight matrices $W_{1,0}, \ldots, W_{\ell_{\max},0} \in \mathbb{R}^{d\times m}$; rank $r < \min(d, m)$; dataset size N; loss $\mathcal{L}$; step size η; minibatch size B; .

Sample minibatch $\mathcal{B} \subset [N]$ with $|\mathcal{B}| = B$ (Poisson rate $q = B/N$)
for each layer $\ell' = 1, \ldots, \ell_{\max}$ **do**
  Sample $A_{\ell'} \sim \mathcal{N}(0,1)^{r\times m}$ and freeze it, Initialize $B_{\ell'}^{(0)} \leftarrow 0 \in \mathbb{R}^{d\times r}$
  $W_\ell \leftarrow W_{\ell,0}$
  $B_{\ell'}^{(1)} \leftarrow B_{\ell'}^0 A_{\ell'} - \eta \nabla_{B_\ell} \mathcal{L}\big(W_\ell, \mathcal{B}\big)$
end for
return $W_\ell \leftarrow W_\ell + B_{\ell'}^{(1)} A_{\ell'}$ for $\ell \in [\ell_{\max}]$.

LoRA enjoys inherent privacy Before proceeding to our formal theoretical results, we first motivate our results using an application. Low-Rank Adaptation (LoRA) [Hu et al., 2022] is one of the most popular parameter-efficient fine-tuning approaches for Large Language Model. LoRA augments a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times m}$ by a *low-rank* update:

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

During fine-tuning, W_0 is frozen; only the small factors (B, A) carry trainable degrees of freedom. This keeps fine-tuning computationally efficient while leaving the base model intact. Algorithm 1 lists the basic steps of the algorithm.

They key observation is that if A is initialized Gaussian and then *frozen* while we update B by

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

then after T steps, we can write

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

Thus each step uses a gradient $\nabla_W \mathcal{L}(W_t)$ that is *right-projected* by the random matrix $A^\top A$ (a rank-r Wishart). This is exactly the projection mechanism we study. Our main result in Theorem 2 shows that W_T enjoys (ε, δ)-DP guarantees where ε, δ depends on dimension of the weight matrices, spectral properties of the gradient matrices, and properties of the dataset collection. While by itself the privacy parameters are not very satisfactory, we believe that the fact that the very popular algorithm already enjoys these guarantees, as has been hinted by previous work [Malekmohammadi and Farnadi, 2025] is already interesting. Additionally, in Section 4 we also discuss algorithmic techniques to amplify these privacy guarantees.

3 Projection mechanism for vector outputs

In this section, we consider the projection mechanism, defined in Definition 1 for vector-valued queries $f : \mathcal{S} \to \mathbb{R}^d$. This setting is a special case of the more general mechanism that projects matrices, but its simpler structure lets us give a clear proof sketch and build intuition for how the mechanism operates. We first establish approximate differential privacy in Section 3.1 and then prove convergence for gradient descent in Theorem 3. Lastly we show how to amplify the privacy of this mechanism by adding small amounts of Gaussian noise.

3.1 Privacy guarantee for the vector case

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

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

When the context is clear, we write $\rho := \rho(\mathcal{D}, f)$ for brevity. 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 1c). Crucially, much like a sensitivity parameter in additive mechanisms ρ is a property of $\mathcal{D}, f$ fixed by the dataset collection and the query and is not controlled by the projection mechanism itself. At the end of this subsection we discuss algorithmic techniques that can increase the effective alignment (Lemma 4).

To recall, for a given r, σ, the projection mechanism samples a Wishart matrix

$$M \coloneqq \sum_{i=1}^{r} z_i z_i^\top = ZZ^\top, \qquad \text{where} \quad z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \sigma^2 I_d\right),\ Z = [z_1, \cdots,\ z_r] \in \mathbb{R}^{d \times r},$$

and releases $Mf(S)$. Here r is the degrees of freedom (equivalently, the rank proxy) of the Wishart distribution and is an algorithmic choice. Increasing r intuitively requires more random bits, as well as increases the rank of M, and improves both ε_ρ and δ_ρ (Figure 1b). Conversely, smaller r can be attractive in practice due to reduced memory and compute at the cost of privacy. Thus, choosing r entails a clear *trade-off between computational efficiency (smaller r) vs.tighter privacy guarantees (larger r).*

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 $f, \mathcal{D}$ as defined in Equation (2). Then, for any $R > (\sqrt{d} + \sqrt{r})^2$, $t > 0$, and $\eta > 0$ the projection mechanism with rank r and variance σ^2 is $(\varepsilon_\rho, \delta_\rho)$-DP on $\mathcal{D}$, where*

$$\delta_\rho = 1 - \bar{\Phi}\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}}\right) \mathbb{P}(\chi_r^2 > t + \eta) + 2\exp\!\left(-c(\sqrt{R} - \sqrt{d} - \sqrt{r})^2\right)$$

$$\varepsilon_\rho \le \frac{|d - r - 1|}{2} \log\left(\rho + \sqrt{1-\rho^2}\sqrt{\frac{R^2}{t^2} - 1}\right) + \frac{R^3\sqrt{2(1-\rho)}}{2t^2}$$

Proof sketch of Theorem 1 Let $v = f(S)$ and $v' = f(S')$ with $\|v\| = \|v'\| = 1$. By Lemma 13, on $\left\{y :\ v^\top y > 0\right\}$ the law of Mv has density

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

and similarly for $p_{v'}(y)$, where $C_{r,d} = \frac{r^{\frac{r+d-1}{2}}}{2^{\frac{r}{2}}(2\pi)^{\frac{d-1}{2}}\Gamma(r/2)}$. For any measurable $\mathcal{Y} \subseteq \operatorname{supp}(Mv) \cap \operatorname{supp}(Mv')$,

$$\frac{\mathbb{P}\left(Mv \in \mathcal{Y}\right)}{\mathbb{P}\left(Mv' \in \mathcal{Y}\right)} \le \sup_{y \in \mathcal{Y}} \frac{C_{r,d}(v^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{r\|y\|^2}{2v^\top y}\right)}{C_{r,d}(v'^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{r\|y\|^2}{2v'^\top y}\right)}$$

To control the above ratio, we keep y away from the boundary of either support and from excessively large norms by defining the good set

$$\mathcal{Y}_{t,R} \coloneqq \left\{ y : \; v^\top y \geq t, \; v'^\top y \geq t, \; \|y\| \leq R \right\}.$$

We define the probability ratio conditioned on the good event as $L_{\max} = \sup_{y \in \mathcal{Y}_{t,R}} \frac{\mathbb{P}(Mv \in \mathcal{Y})}{\mathbb{P}(Mv' \in \mathcal{Y})}$. Then, for any $\mathcal{Y} \subseteq \operatorname{supp}(Mv)$,

$$\begin{aligned}\mathbb{P}\left(Mv \in \mathcal{Y}\right) &= \mathbb{P}\left(Mv \in \mathcal{Y} \cap \mathcal{Y}_{t,R}\right) + \mathbb{P}\left(Mv \in \mathcal{Y} \cap \mathcal{Y}_{t,R}^c\right) \\ &\leq L_{\max}\mathbb{P}\left(Mv' \in \mathcal{Y} \cap \mathcal{Y}_{t,R}\right) + \mathbb{P}\left(Mv \in \mathcal{Y}_{t,R}^c\right) \leq L_{\max}\mathbb{P}\left(Mv' \in \mathcal{Y}\right) + \mathbb{P}\left(\mathcal{Y}_{t,R}^c\right)\end{aligned}$$

and symmetrically with v, v' swapped. Bounding $L_{\max}$ by decomposing $v' = \rho v + \sqrt{1-\rho^2}\, w$ with $w \perp v$ yields the stated ε_ρ.

Next, we bound the failure probability of $\mathcal{Y}_{t,R}$ with δ_ρ. While the proof for this is slightly more involved, it starts from the observation that $X \coloneqq v^\top M v / \sigma^2 \sim \chi^2_r$ and conditional on $X = x$, $v^\top M v' \mid X = x$ is distributed as a Gaussian. Using these, we lower bound $\Pr\left(v^\top M v' \geq t \;\middle|\; X = x\right)$ with the product of the χ^2_r and $\mathcal{N}$ tails. Second, the operator-norm event $\{\|M\| > R\}$ is controlled by $\|M\| = \sigma^2 \|G\|^2$ where G is a standard Gaussian random matrix. Combining this yields the required bound on δ_ρ. □

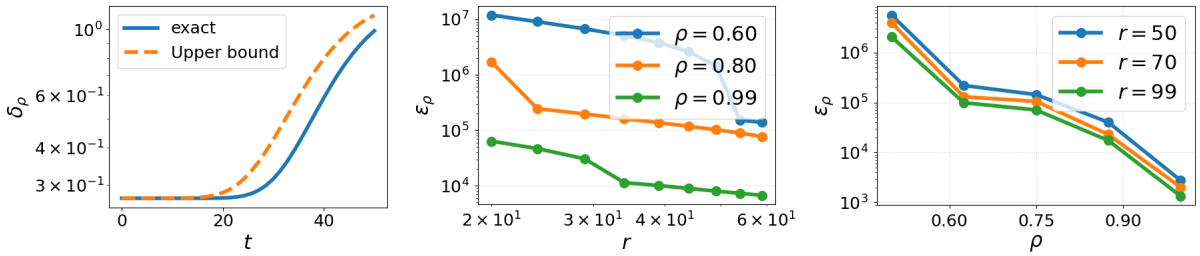
On the dependence of ρ on n. Intuitively, ρ captures how stable the direction of the normalised query is under a record change. For average-like queries, a single replacement contributes an $O\left(1/n\right)$ perturbation to the unnormalised vector, which in turn yields $\rho \approx 1 - \widetilde{O}\left(1/n^2\right)$. Concretely, let $\tilde{f}(S) = \frac{1}{n}\sum_{i=1}^n g(x_i)$ with $\|g(x)\| \leq L$ and assume $\inf_{S \in \mathcal{D}} \left\|\tilde{f}(S)\right\| \geq c_0 > 0$, for a constant c_0. Then neighbouring datasets S, S' satisfy $\left\|\tilde{f}(S) - \tilde{f}(S')\right\| \leq 2L/n$, which implies $\|f(S) - f(S')\| \leq \frac{4L}{c_0 n}$. Since $1 - u^\top v = \frac{1}{2}\|u - v\|^2$ for unit u, v, we obtain $\rho \geq 1 - \frac{1}{2}\|f(S) - f(S')\|^2 \geq 1 - \frac{8L^2}{c_0^2 n^2}$.

Additional analysis parameters t, η, R. The statement of Theorem 1 introduces $t > 0$, $\eta \geq 0$, and $R > (\sqrt{d} + \sqrt{r})^2$. These are analysis parameters used to obtain a tight yet concise bound; they are *not* hyper-parameters of the mechanism. In practice one evaluates the bound numerically by optimising over these quantities. The slack η decouples a χ^2_r tail from a (conditional) Gaussian tail in the proof of δ_ρ, yielding a closed-form expression that is conservative; empirically, the exact mixed-tail calculation can be smaller (see Figure 1a). The parameters R and t define the good set and serve as knobs between δ_ρ and ε_ρ: increasing t leads to a monotonic decrease in ε_ρ, while simultaneously increasing the bound on δ_ρ. The parameter R controls the spectral-concentration event for the Wishart matrix: larger R strengthens concentration and *decreases* δ_ρ, but *increases* ε_ρ. A practical choice is to take $R = \left(\sqrt{d} + \sqrt{r} + u\right)^2$ with a small $u > 0$, which makes the spectral term in δ_ρ negligible once r is moderate. Given a target $\bar{\delta} \in (0, 1)$ and a maximum rank $r_{\max} \ll d$, one can solve the δ_ρ bound for the largest admissible t and then plug that t into the ε_ρ bound, leading to the following corollary.

Corollary 1. *Let ρ be the minimum alignment as defined in Equation (2). Then for any $0 < r \leq d$, $\delta > 2e^{-cr}$ (for an absolute constant $c > 0$ from Gaussian spectral concentration), the projection mechanism is $(\varepsilon_\rho, \delta)$-DP with*

$$\varepsilon_\rho \leq \tilde{C}\left(\frac{d - r + 1}{2}\log\left(\rho + \sqrt{(1-\rho^2)\left(\tfrac{d^2}{\rho^2 r^2} - 1\right)}\right) + \frac{d^3\sqrt{2(1-\rho)}}{2\rho^2 r^2}\right),$$

for a universal constant $\tilde{C} > 0$.



(a) (b) (c)

Figure 1: 1a shows exact δ via finite integration vs our bound,1b shows ε_ρ as a function of r, and 1c shows ε_ρ as a function of ρ. In figures 1b and 1c, ε_ρ was computed by doing a sweep over all analysis parameters to find the best parameter with $\delta < 0.1$.

When $\rho = 1$ we have $f(S) = f(S')$, so the induced output distributions coincide and the true privacy loss satisfies $\varepsilon_\rho = 0$, which is also reflected in our bound. We note however that his intuition cannot be observed in our δ_ρ bound, an artifact of the conservative decoupling and concentration steps used in the analysis. See Figure 1a for an illustration of this gap.

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

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

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

Lemma 4. *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).

3.2 Applications

In this section, we highlight three potential applications of the projection mechanism for the case. In Section 4, we highlight our main application to show that LoRA is inherently private.

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. In Theorem 3,we provide a convergence guarantee and identify regimes in which RP–GD improves upon DP–GD when the dataset collection exhibits well-aligned gradient sums. For more details see Section E.

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.

4 Projection mechanism for matrix outputs

In this section we consider the projection mechanism applied to $f : \mathcal{S} \to \mathbb{R}^{d\times m}$ with matrix-valued output space ($m > 1$). We prove a privacy guarantee (Section 4.1) and apply this result to the popular finetuning algorithm, LoRA, showing that LoRA inherits the same guarantee (Section 2).

4.1 Privacy guarantee

Similar to the vector case, to obtain meaningful privacy guarantees for the projection mechanism, we assume the outputs of the query f on any neighboring datasets $S, S' \in \mathcal{D}$ are sufficiently aligned (equation A1). We also assume that the outputs have rank k and bounded spectral norm (equation A2). As in the vector-output setting, these assumptions depend only on the dataset collection and the query function. At the end of this section, we provide algorithmic solutions to amplify these values.

We formally define the assumptions as follows. For matrix outputs in $\mathbb{R}^{d\times m}$ with $d \geq m$ and $\operatorname{rank}(f(S)) \leq k \leq \min(d, m)$, we measure alignment by the cosine of the k-th principal angle between the *column spaces* of $f(S)$ and $f(S')$:

$$\rho_k(f, \mathcal{D}) = \min_{S \overset{H}{\sim} S',\; S,S' \in \mathcal{D}} \sigma_k\left(U_f(S)^\top U_f(S')\right), \tag{A1}$$

where $U_f(S)$, $U_f(S')$ have orthonormal columns spanning $\operatorname{Col}(f(S))$ and $\operatorname{Col}(f(S'))$.

To ensure the outputs have rank k and bounded norm, we impose: there exist constants $\xi_k > 0$ and $\xi_1 < \infty$ such that for all $S \in \mathcal{D}$,

$$\sigma_k\big(f(S)\big) \geq \xi_k \quad \text{and} \quad \sigma_1\big(f(S)\big) \leq \xi_1 \tag{A2}$$

where $\sigma_i(\cdot)$ denotes the i-th largest singular value.

Projection mechanism. To recall, the mechanism releases MV where $V = f(S) \in \mathbb{R}^{d\times m}$ and $M = \sum_{i=1}^{r} z_i z_i^\top$ with $z_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_d/r)$. This is a rank-$r$ Wishart projection.

Theorem 2. *For a dataset collection $\mathcal{D}$ and a query function f with outputs in $\mathbb{R}^{d\times m}$ of rank k, let ρ_k be the (column-space) alignment as in equation A1, and let ξ_k, ξ_1 be as in equation A2. Further, let $\Delta = \max_{S\overset{H}{\sim}S',\ S,S'\in\mathcal{D}} \|f(S) - f(S')\|_2$. Then for any tuning parameters $t > 0$, and $R > \xi_1(\sqrt{d}+\sqrt{r})^2$, the rank-$r$ projection mechanism with $M = \sum_{i=1}^{r} z_i z_i^\top$ and $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d/r)$ is (ε, δ)-DP with*

$$\delta \le e^{-\frac{1}{2}\left(\sqrt{k}-\sqrt{r}-(\sqrt{t}+\eta)/\xi_k\right)_+^2} + 2e^{-\frac{1}{2}\left(\frac{(\rho_k-1)t}{\sqrt{1-\rho_k^2}R}-\sqrt{m}-\sqrt{r}\right)_+^2} + 2e^{-c\left(\sqrt{\frac{R}{\xi_1}}-\sqrt{d}-\sqrt{r}\right)_+^2},$$

$$\varepsilon \le \frac{(m-r)k}{2}\log\left(1+\frac{2\xi_1}{\xi_k^2}\,\Delta\right) + \frac{r(d-r+1)}{2}\log\left(1+\frac{R}{t}\,\Delta\right) + \frac{r^{3/2}\sqrt{k}\;R^3}{\sqrt{2}\,t^2}\Delta.$$

On the dependence of ρ_k on dataset size n Similar to the vector case, we can show that the alignment parameter ρ_k increases with n. Let $\tilde{f}(S) = \frac{1}{n}\sum_{i=1}^{n} G(x_i) \in \mathbb{R}^{d\times m}$ with $\|G(x)\|_2 \le L$; the sensitivity of $\tilde{f}$ is $\left\|\tilde{f}(S) - \tilde{f}(S')\right\| \le \frac{2L}{n}$. Under Equation (A2) and the rank assumption of Theorem 2, let $\theta_{\max}$ be the largest principal angle between the k-dimensional column spaces of $\tilde{f}(S)$ and $\tilde{f}(S')$. Using the $\sin\Theta$ theorem for column subspaces, we obtain $\sin(\theta_{\max}) = \|\sin\Theta\| \le \frac{\|\tilde{f}(S)-\tilde{f}(S')\|}{\xi_k} \le \frac{2L}{n\xi_k}$. Using $\rho_k = \cos\theta_{\max}$ and $1-\cos\theta \le \sin^2\theta$ for $\theta \in [0, \pi/2]$, we have that $\rho_k \ge 1 - \frac{4L^2}{n^2\xi_k^2}$. Thus for average-like matrix queries, $\rho_k \approx 1 - \widetilde{O}(1/n^2)$. In our matrix projection theorem, larger ρ_k directly reduces the second term in δ.

4.2 Application to LoRA

As discussed before, a natural application of our privacy guarantee for the low-rank projection mechanism is *Low-Rank Adaptation (LoRA)*[Hu et al., 2022]. Concretely, let $G_t = \nabla_W \mathcal{L}(W_t)$. Hao et al. [2024] shows that LoRA update with fixed A is

$$B_{t+1} = B_t - \eta\, G_t A^\top \qquad \Longrightarrow \qquad W_T = W_0 + B_T A = W_0 - \eta \sum_{t=1}^{T} G_t (A^\top A).$$

Thus LoRA performs gradient descent with the gradient *right-projected* by the random Wishart matrix $A^\top A$, which is precisely a projection mechanism on a matrix-valued output. Consequently, LoRA inherits our privacy guarantee (Theorem 2) automatically. In practice, LoRA is applied across multiple layers and trained with mini-batches. The privacy guarantee in Theorem 2 can be simply composed over multiple layers and amplified via minibatch subsampling as is common in DP-SGD.

Algorithmic techniques to improve LoRA's privacy guarantees Analogous to the RP-GD analysis in Theorem 3, applying LoRA with a single initialization of A requires Assumption A1 and A2 to hold for the cumulative gradient trajectory $\sum_{t=1}^{T} G_t$ under any neighbouring datasets in the collection, since early misalignment can be amplified over the course of optimization. A practical relaxation is to impose these assumptions per step (rather than on the sum) and *resample* A every τ steps, incurring an additional composition over restarts. We include this resampling trick in Algorithm 2, which is in the same spirit as Hao et al. [2024].

Algorithm 2 LoRA with resampled A on one layer

Input: pretrained model parameters $W_0 \in \mathbb{R}^{d\times m}$, rank r with $r < \min(d, m)$, input $X \in \mathbb{R}^{N\times k\times n}$, loss function $\mathcal{L}$, number of rounds T and τ, step size η, mini-batch size B

for $t = 0, \dots T-1$ **do**
  Initialize $B_t^{(0)} \in \mathbb{R}^{d\times r}$ and $A_t \in \mathbb{R}^{r\times k}$ randomly ▷ resample A
  for $j = 1, \dots \tau$ **do**
    Random sample an example $x \in \mathbb{R}^{k\times n}$ from the dataset X with probability B/N
    $W_t^j \leftarrow W_t + B_t^{(j-1)} A_t$ (should it be ▷ update model
    $y_j \leftarrow W_t^j x$ ▷ evaluate
    $B_t^{(j)} \leftarrow B_t^{(j-1)} - \eta \frac{\partial \mathcal{L}(y)}{\partial B}|_{y=y_j}$ ▷ update B
  end for
  $W_{t+1} = W_t + B_t^{(\tau)} A_t$
end for
return W_T

Role of ξ_k and ξ_1. Further, it is also possible to improve the spectral properties of the gradient. Assumption A2 requires uniform spectral bounds over $\mathcal{D}$: $\sigma_k(f(S)) \geq \xi_k > 0$ and $\sigma_1(f(S)) \leq \xi_1 < \infty$ for all $S \in \mathcal{D}$. Concretely, ξ_1 can be reduced by enforcing spectral regularisation on the gradients e.g. using spectral normalization of linear layers, gradient clipping in operator norm, or whitening that rescales dominant directions. Conversely, ξ_k can be increased by adding a small regulariser term to the loss e.g. ridge regulariser injects a λW term into the gradient and lifts smaller singular values of the gradient, curvature damping adds λI to the gradient to prevent rank collapse; both of these techniques raise the floor on ξ_k. In our bounds this improves $\sqrt{t}/\xi_k$ in δ and reduces the factor $(m-r)k \log\left(1 + \frac{2\xi_1}{\xi_k^2}\Delta\right)$ in ε. We leave the exploration of further implications of these techniques to future work.

5 Discussion and open questions

The main contribution of our work is showing that the random projection mechanism is inherently differentially private. Using this analysis we are able to show that LoRA also enjoys certain privacy guarntees. We also show how to improve this privacy guarantee algorithmically, either by adding a small amount of additive gaussian noise or via algorithmic techniques like regularisation and smoothening. Our proof relies on new technical tools in analysing the likelihood ratio of more complex distributions than the classical distributions analyses in Differential Privacy.

However, there are several opportunities for further work. First, the privacy parameter is complicated and hard to tune, additionally it scales with properties of the dataset like the alignment. We believe techniques like smooth sensitivity, inverse-sensitivity mechanism, and PTR may be helpful in overcoming this issue. Second, we describe several algorithmic techniques that can improve the privacy guarantee and we think it will be interesting to consider their impact on the utility of the algorithm. Finally, while we list a few applications of the projection mechanism, it would be interesting to find other applications of this mechanism.

Related works Random projection has been 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 random projection primarily for dimensionality reduction, improving the privacy–utility trade-off by removing dimension dependence from convergence guarantees in many private algorithms [Jiang et al., 2025, Kasiviswanathan, 2021, Li et al., Sheffet, 2019]. By contrast, far fewer works explicitly leverage the inherent randomness of the projection itself as a source of privacy.

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, SignSGD amplifies privacy guarantees. Perhaps most related to our work, Hao et al. [2024], Malekmohammadi and Farnadi [2025] argue that LoRA can be viewed as gradient compression through low-dimensional random projection and induces training dynamics resembling DP-SGD for certain architectures. However, to our knowledge, we are the first to provide a formal differential privacy guarantee for compression achieved via random projection for arbitrary model choice.

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.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *Proceedings of the 2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, FOCS '14, page 464–473, 2014.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. *URL https://arxiv. org/abs/2305.14314*, 2, 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.

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. *arXiv preprint arXiv:2402.12354*, 2024.

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 *International Conference on Learning Representations*, 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. URL `https://openreview.net/forum?id=u6OSRdkAwl`. 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. URL `https://openreview.net/forum?id=thCqMz1ZXw`.

Shiva Prasad Kasiviswanathan. Sgd with low-dimensional gradients with applications to private and distributed learning. In Cassio de Campos and Marloes H. Maathuis, editors, *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *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.

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. *CoRR*.

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. *arXiv preprint arXiv:2310.08659*, 2023.

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, 2025.

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.

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. *arXiv preprint arXiv:2403.12313*, 2024.

Roman Vershynin. *High-dimensional probability.* 2018.

Per-Åke Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2):217–232, 1973.

Hermann Weyl. Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, 71(4):441–479, 1912.

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*, 2023.

Appendix

A Mathematical Preliminaries

Definition 6. *Matrices A and B are similar if there exists a matrix W so that*

$$A = WBW^{-1}$$

Lemma 5. *Let A and B be symmetric matrices (of the same dimension). If A and B are similar matrices, then they have the same eigenvalues.*

Proof. Let A and B be similar matrices, this means there exists W so that $A = WBW^{-1}$. Let x be an eigenvector of A with λ its corresponding eigenvalue. This means

$$Ax = WBW^{-1}x = \lambda x$$

by right multiplicating with W^{-1} we see that this is equivalent to

$$B(W^{-1}x) = \lambda(W^{-1}x).$$

□

Lemma 6. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix, and $Q \in \mathbb{R}^{m\times m}, Q' \in \mathbb{R}^{n\times n}$ orthogonal matrices then A and $A' = QAQ'$ have the same singular values.*

Proof. Let $A' = QAQ'$, then

$$A'^{\top}A' = (Q'^{\top}A^{\top}Q^{\top})(QAQ') = Q'^{\top}A^{\top}AQ'.$$

By Lemma 5 we know $A'^{\top}A'$ and $A^{\top}A$ have the same eigenvalues, which finishes the proof. □

Lemma 7 (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 8. *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 9 (Theorem 3.4.5 [Vershynin, 2018])**.** *Let u be a random vector uniformly distributed on the unit sphere in $\mathbb{R}^d$ (or $w \sim \mathcal{N}(0, I_d)$ and let $u = \frac{w}{\|w\|}$). Then for any unit vector $v \in \mathbb{R}^d$ and $t > 0$, we have*

$$\mathbb{P}\left[\langle u, v\rangle \geq t\right] \leq 2e^{-\frac{t^2 d}{2}}.$$

Lemma 10 (Perturbation bounds for pseudo-inverse, Theorem 4.1 in [Wedin, 1973])**.** *Let $A \in \mathbb{R}^{m\times n}$ of rank r, $B \in \mathbb{R}^{m\times n}$ of rank s, then*

$$\left\|B^{\dagger} - A^{\dagger}\right\|_F \leq \sqrt{2}\max\left\{\left\|A^{\dagger}\right\|_2^2, \left\|B^{\dagger}\right\|_2^2\right\}\|B - A\|_F$$

Lemma 11 (Weyl's inequality, [Weyl, 1912])**.** *Let $\Delta \in \mathbb{R}^{m\times n}$ be a perturbation of arbitrary magnitude. Denote $\tilde{X} = X + \Delta$ with singular values $\tilde{\sigma}_1 \geq \tilde{\sigma}_2 \geq \cdots \geq \tilde{\sigma}_n \geq 0$. Then,*

$$|\tilde{\sigma}_i - \sigma_i| \leq \|\Delta\|_2, \text{ for } i = 1, \ldots, n$$

Lemma 12. *Let $W \sim Wishart_d(r, I_d\sigma^2)$, then for any $g \in \mathbb{R}^d$ with $\|g\|_2 = 1$,*

$$\mathbb{E}g^T WW^T g = \sigma^4 r(r + d + 1)$$

Proof. We know that for a wishart distribution W,

$$\mathbb{E}W = r\sigma^2 I_d, \quad \mathrm{Cov}(W_{ij}) = n(v_{ij}^2 + v_{ii}v_{jj}).$$

Let $\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \ldots \\ 1 \end{pmatrix} \in \mathbb{R}^d$. Then,

$$\begin{aligned} \mathbb{E}g^T WW^\top g &= g^\top \left(\mathrm{Cov}(W) + \mathbb{E}(W)\mathbb{E}(W)^\top \right) g \\ &= g^\top \left(\sigma^4 r(I_d + \mathbf{1}\mathbf{1}^\top) + \sigma^4 r I_d \right) g \\ &= \sigma^4 r(r + d + 1) \end{aligned} \tag{3}$$

□

B Privacy of Projection Mechanism on vectors

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 $f, \mathcal{D}$ as defined in Equation (2). Then, for any $R > (\sqrt{d} + \sqrt{r})^2$, $t > 0$, and $\eta > 0$ the projection mechanism with rank r and variance σ^2 is $(\varepsilon_\rho, \delta_\rho)$-DP on $\mathcal{D}$, where*

$$\delta_\rho = 1 - \bar{\Phi}\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}} \right) \mathbb{P}(\chi_r^2 > t + \eta) + 2\exp\left(-c(\sqrt{R} - \sqrt{d} - \sqrt{r})^2\right)$$

$$\varepsilon_\rho \leq \frac{|d - r - 1|}{2} \log\left(\rho + \sqrt{1-\rho^2}\sqrt{\frac{R^2}{t^2} - 1} \right) + \frac{R^3\sqrt{2(1-\rho)}}{2t^2}$$

Proof. By Lemma 13 we know that for y so that $v^\top y > 0$

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

So if we assume $y \in \mathcal{Y}_{t,R}$ with

$$\mathcal{Y}_{t,R} := \{Y : v^\top y > t, v'^\top y > t, \|y\| \leq R\}$$

then for any subset $\mathcal{Y} \in \mathcal{Y}_{t,R}$,

$$\begin{aligned} \frac{\mathbb{P}(Mv \in \mathcal{Y})}{\mathbb{P}(Mv' \in \mathcal{Y})} &= \frac{\int_{y\in\mathcal{Y}} C_{r,d,\sigma}(v^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2 v^\top y}\right) dy}{\int_{y\in\mathcal{Y}} C_{r,d,\sigma}(v'^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2 v'^\top y}\right) dy} \\ &\leq \sup_{y\in\mathcal{Y}} \frac{C_{r,d}(v^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{r\|y\|^2}{2v^\top y}\right)}{C_{r,d}(v'^\top y)^{\frac{r-d-1}{2}} \exp\left(-\frac{r\|y\|^2}{2v'^\top y}\right)} := L_{\max} \end{aligned} \tag{4}$$

For arbitrary $\mathcal{Y} \subseteq \text{Support(Mv)}$,

$$\begin{aligned}\mathbb{P}(Mv \in \mathcal{Y}) &= \mathbb{P}(Mv \in \mathcal{Y} \cap \mathcal{Y}_{t,R}) + \mathbb{P}(Mv \in \mathcal{Y} \cap \mathcal{Y}_{t,R}^c) \\ &\overset{(a)}{\leq} L_{\max}\mathbb{P}(Mv' \in \mathcal{Y} \cap \mathcal{Y}_{t,R}) + \mathbb{P}(Mv \in \mathcal{Y}_{t,R}^c) \\ &\overset{(b)}{\leq} L_{\max}\mathbb{P}(Mv' \in \mathcal{Y}) + \mathbb{P}(Mv \in \mathcal{Y}_{t,R}^c) \overset{(c)}{\leq} L_{\max}\mathbb{P}(Mv' \in \mathcal{Y}) + \delta\end{aligned} \tag{5}$$

where step (a) follows from Equation (4) together with $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$, step (b) again uses $\mathbb{P}(A \cap B) \leq \min(\mathbb{P}(A), \mathbb{P}(B))$, and step (c) follows from Equation (6).

The same argument holds for the other direction, with $\mathbb{P}(Mv' \in \mathcal{Y}) \leq L_{\max}\mathbb{P}(Mv \in \mathcal{Y}) + \delta'$ for any $\mathcal{Y}$ where $\delta' = \mathbb{P}(Mv' \in \mathcal{Y}_{t,R}^c)$.

So it remains to show that $\max\{\delta, \delta'\} \leq \delta_\rho$. For $M = \sum_{i=1}^r z_i z_i^\top$ where $z_i's$ are i.i.d. $\mathcal{N}(0, \sigma^2 I_d)$, let $Y \overset{d}{=} Mv$.

$$\mathbb{P}(Y \in \mathcal{Y}_{t,R}) = \mathbb{P}(Mv \in \mathcal{Y}_{t,R}) = \mathbb{P}_M\left[v^\top Mv > t, v'^\top Mv \geq t, \|Mv\| \leq R\right] \tag{6}$$

equivalently for $Y' \overset{d}{=} Mv'$

$$\mathbb{P}(Y' \in \mathcal{Y}_{t,R}) = \mathbb{P}(Mv' \in \mathcal{Y}_{t,R}) = \mathbb{P}_M\left[v^\top Mv' > t, v'^\top Mv' \geq t, \left\|Mv'\right\| \leq R\right] \tag{7}$$

As v and v' are both unit vectors these two events can $\|Mv'\| \leq R$ and $\|Mv\| \leq R$ are both implied by $\|M\| \leq R$. So both $\mathbb{P}(Y \in \mathcal{Y}_{t,R})$ and $\mathbb{P}(Y' \in \mathcal{Y}_{t,R})$ are lower bounded by

$$\mathbb{P}\left[v^\top Mv' > t, v'^\top Mv' \geq t, \|M\| \leq R\right] = \mathbb{P}\left[v^\top Mv' > t, v^\top Mv \geq t, \|M\| \leq R\right]$$

We will first focus on the event

$$\mathcal{G}_t = \{v^\top Mv' > t, v^\top Mv \geq t\}$$

(note that $\{v^\top Mv' > t, v'^\top Mv' \geq t\}$ occurs with the same probability as $\{v^\top Mv' > t, v^\top Mv \geq t\}$ due to the randomness being over M and v and v' both being unit vectors). In order to bound this let us define

$$\begin{aligned}g_i &:= \frac{1}{\sigma} v^\top z_i \\ h_i &:= \frac{1}{\sigma} v'^\top z_i\end{aligned}$$

both of these are uni variate standard random normal variables and

$$\sum_{i=1}^r g_i^2 = \frac{1}{\sigma^2}\sum_{i=1}(v^\top z_i)^2 = \frac{1}{\sigma^2}\sum_{i=1}^r v^\top z_i z_i^\top v = \frac{1}{\sigma^2} \cdot v^\top Mv$$

as well as

$$g^\top \cdot h = \sum_{i=1}^r g_i h_i = \frac{1}{\sigma^2}\sum_{i=1}^r v^\top z_i z_i^\top v' = \frac{1}{\sigma^2} \cdot v^\top Mv'$$

where $g = (g_1, \ldots, g_r)$ and $h = (h_1, \ldots, h_r)$. From now on we will denote $X := \frac{1}{\sigma^2} \cdot v^\top Mv$, and $Z := v^\top Mv'$ for simplicity of notation. As the g_i are univariate standard normal variables, we have

$X \sim \chi^2_r$. Next, we would like to partition h_i into a part independent and a part dependent on g_i. For this, we define the unit vector

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

Then, we decompose v' into $v' = \rho v + \sqrt{1-\rho^2} w$, which gives us

$$h_i := \frac{1}{\sigma}(\rho v^\top z_i + \sqrt{1-\rho^2} w^\top z_i) = \rho g_i + \sqrt{1-\rho^2}\varepsilon_i$$

where $\varepsilon_i := w^\top z_i \sim \mathcal{N}(0, \sigma^2)$. Therefore,

$$\frac{1}{\sigma^2} \cdot Z = g^\top \cdot h = \sum_{i=1}^{r} \rho g_i^2 + \sqrt{1-\rho^2} g_i \varepsilon_i$$

Here, we constructed ε in such a way that ε_i is independent of g_i as

$$\mathrm{Cov}(g_i, \varepsilon_i) = \mathbb{E}[g_i \varepsilon_i] = v^\top \mathbb{E}[z_i z_i^\top] w = \frac{1}{\sigma^2} v^\top w = 0.$$

This allows us to condition on g_i and obtain $Z|X = x \sim \mathcal{N}(\sigma^2 \rho \cdot x, \sigma^4(1-\rho^2)x)$. Now, we can lower bound the probability of the event $G_t := \{v^\top M v' > t, v^\top M v \geq t\}$,

$$\begin{aligned} \mathbb{P}(v^\top M v > t, v'^\top M v > t) &= \mathbb{P}(X > t/\sigma^2, Z > t) \\ &= \mathbb{E}_{X,Z}[\mathbf{1}_{X > t/\sigma^2} \cdot \mathbf{1}_{Z>t}] \\ &= \mathbb{E}_X[\mathbf{1}_{X>t/\sigma^2} \mathbb{E}_Z[\mathbf{1}_{Z>t}|X]] \end{aligned}$$

Next, we will lower bound $\mathbb{E}_Z[\mathbf{1}_{Z>t}|X]$ by noting that

$$\mathbb{P}(Z > t | X = x) = \bar{\Phi}\left(\frac{t - \rho x \sigma^2}{\sigma^2 \sqrt{(1-\rho^2)x}}\right),$$

so

$$\mathbb{E}_X[\mathbf{1}_{X>rt} \mathbb{E}_Z[\mathbf{1}_{Z>t}|X]] = \mathbb{E}_X\left[\mathbf{1}_{X>rt} \bar{\Phi}\left(\frac{t - \rho X \sigma^2}{\sigma^2 \sqrt{(1-\rho^2)X}}\right)\right]$$

We further observe that $f(x) := \frac{t - \rho x \sigma^2}{\sigma^2 \sqrt{(1-\rho^2)x}}$ is decreasing in x as

$$\frac{d}{dx} f(x) = -\frac{t + \rho \sigma^2 x}{2\sigma^2 \sqrt{1-\rho^2} x^{3/2}},$$

for $\rho \geq 0$. Hence, for $\rho, t, \sigma > 0$, $\bar{\Phi}\left(\frac{t - \rho x \sigma^2}{\sigma^2 \sqrt{(1-\rho^2)x}}\right)$ is increasing in x, and for any $x^\star > rt$

$$\mathbb{P}(\mathcal{G}_t) \geq \bar{\Phi}\left(\frac{t - \rho x^\star \sigma^2}{\sigma^2 \sqrt{(1-\rho^2)x^\star}}\right) \mathbb{P}(\chi^2_r > x^\star).$$

We want to choose $x^\star$ to maximize this lower bound. Let's choose it as $x^\star = t/\sigma^2 + \eta$ for any $\eta \geq 0$. Then,

$$\mathbb{P}(\mathcal{G}_t) \geq \bar{\Phi}\left(\frac{t/\sigma^2(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t/\sigma^2+\eta)}}\right)\mathbb{P}(\chi_r^2 > t/\sigma^2 + \eta)$$

We further define $\tilde{E}_R = \left\{z_i : \left\|\sum_{i=1}^r z_i z_i^\top\right\| \leq R\right\}$ and $E_R = \{z_i : \left\|\sum_{i=1}^r z_i z_i^\top v\right\| \leq R\}$. Then, $\tilde{E}_R \subset E_R$ and $\mathbb{P}(E_R) \geq \mathbb{P}(\tilde{E}_R)$. As $M = \sum_{i=1}^r z_i z_i^\top$ can be decomposed as $\left(\frac{1}{\sqrt{r}}Z\right)\left(\frac{1}{\sqrt{r}}Z\right)^\top$ where $Z \in \mathbb{R}^{d\times r}$ with i.i.d. entries $\mathcal{N}(0,1)$, we can write

$$\|M\| = \left\|ZZ^\top\right\|\sigma^2.$$

Then, we apply Lemma 7 with $t' = \sqrt{R}/\sigma - (\sqrt{d}+\sqrt{r}) > 0$ to get an upper bound on the bad event E_R^c,

$$\mathbb{P}(\tilde{E}_R^c) = \mathbb{P}(\|M\| \geq R) = \mathbb{P}\left(\|Z\| \geq \frac{1}{\sigma}\sqrt{R}\right) \leq 2\exp\left(-c\left(\frac{\sqrt{R}}{\sigma} - \sqrt{d} - \sqrt{r}\right)^2\right) \tag{8}$$

Note this gives us the condition $R \geq \sigma^2(\sqrt{d}+\sqrt{r})^2$. Finally using

$$\mathbb{P}(G_t^c \cup E_R^c) \leq \mathbb{P}(G_t^c) + \mathbb{P}(E_R^c) = 1 - \mathbb{P}(G_t) + \mathbb{P}(E_R^c)$$

we get that

$$\delta_\rho \leq 1 - \bar{\Phi}\left(\frac{t/\sigma^2(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t/\sigma^2+\eta)}}\right)\mathbb{P}(\chi_r^2 > t/\sigma^2 + \eta) + 2\exp\left(-c\left(\frac{\sqrt{R}}{\sigma} - \sqrt{d} - \sqrt{r}\right)^2\right)$$

Finally, we derive the upper bound on $L_{\max}$. For $y \in G_t \cup E_R$, by Lemma 13,

$$\frac{P(y)}{Q(y)} = \left(\underbrace{\frac{v^\top y}{(v')^\top y}}_{A}\right)^{\frac{r-d-1}{2}} \exp\left(\underbrace{\frac{\|y\|^2}{2\sigma^2 y^\top v'} - \frac{\|y\|^2}{2\sigma^2 y^\top v}}_{B}\right) \tag{9}$$

where $U = [u_1, \ldots, u_{d-1}] \in \mathbb{R}^{d\times(d-1)}$ with $\{u_1, \ldots, u_{d-1}\}$ is an orthonormal basis of $v^\perp$.

$$\begin{aligned}
A = \frac{v'^\top y}{v^\top y} &= \frac{\rho v^\top y + \sqrt{1-\rho^2}w^\top P_\perp y}{v^\top y} \\
&= \rho + \sqrt{1-\rho^2}\frac{w^\top P_\perp y}{v^\top y} \leq \rho + \sqrt{1-\rho^2}\frac{\|P_\perp y\|}{v^\top y} \\
&\leq \rho + \sqrt{1-\rho^2}\frac{\sqrt{R^2-t^2}}{t}
\end{aligned}$$

We can bound the exponent (part B) by,

$$\begin{aligned}
B &= \frac{\|y\|^2}{2\sigma^2 y^\top v'} - \frac{\|y\|^2}{2\sigma^2 y^\top v} \\
&= \frac{\|y\|^2 (y)^\top (v - v')}{2\sigma^2 y^\top v' v^\top y} \\
&\leq \frac{R^3\sqrt{2(1-\rho)}}{2\sigma^2 t^2}
\end{aligned}$$

All together this gives

$$\delta_\rho \le 1 - \bar{\Phi}\left(\frac{t/\sigma^2(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t/\sigma^2+\eta)}}\right)\mathbb{P}(\chi_r^2 > t/\sigma^2 + \eta) + 2\exp\left(-c\left(\frac{\sqrt{R}}{\sigma} - \sqrt{d} - \sqrt{r}\right)^2\right)$$

and

$$\varepsilon_\rho \le \frac{r-d-1}{2}\log\left(\rho + \sqrt{1-\rho^2}\frac{\sqrt{R^2-t^2}}{t}\right) + \frac{R^3\sqrt{2(1-\rho)}}{2\sigma^2 t^2}$$

setting $\tilde{t} = t/\sigma^2$ and $\tilde{R} = R/\sigma^2$, we get the bounds:

$$\delta_\rho \le 1 - \bar{\Phi}\left(\frac{\tilde{t}(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(\tilde{t}+\eta)}}\right)\mathbb{P}(\chi_r^2 > \tilde{t} + \eta) + 2\exp\left(-c\left(\sqrt{\tilde{R}} - \sqrt{d} - \sqrt{r}\right)^2\right)$$

and

$$\varepsilon_\rho \le \frac{r-d-1}{2}\log\left(\rho + \sqrt{1-\rho^2}\frac{\sqrt{\tilde{R}^2\sigma^4 - \tilde{t}^2\sigma^4}}{\tilde{t}\sigma^2}\right) + \frac{\tilde{R}^3\sigma^6\sqrt{2(1-\rho)}}{2\sigma^6\tilde{t}^2}$$

with the condition $\tilde{t} > 0$ and $\tilde{R} \ge (\sqrt{d}+\sqrt{r})^2$, which simplifies to (with renaming the variables again)

$$\delta_\rho \le 1 - \bar{\Phi}\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}}\right)\mathbb{P}(\chi_r^2 > t + \eta) + 2\exp\left(-c\left(\sqrt{R} - \sqrt{d} - \sqrt{r}\right)^2\right)$$

$$\varepsilon_\rho \le \frac{r-d-1}{2}\log\left(\rho + \sqrt{1-\rho^2}\frac{\sqrt{R^2-t^2}}{t}\right) + \frac{R^3\sqrt{2(1-\rho)}}{2t^2}$$

□

Corollary 1. *Let ρ be the minimum alignment as defined in Equation (2). Then for any $0 < r \le d$, $\delta > 2e^{-cr}$ (for an absolute constant $c > 0$ from Gaussian spectral concentration), the projection mechanism is $(\varepsilon_\rho, \delta)$-DP with*

$$\varepsilon_\rho \le \tilde{C}\left(\frac{d-r+1}{2}\log\left(\rho + \sqrt{(1-\rho^2}\left(\tfrac{d^2}{\rho^2 r^2} - 1\right)\right) + \frac{d^3\sqrt{2(1-\rho)}}{2\rho^2 r^2}\right),$$

for a universal constant $\tilde{C} > 0$.

Proof. With $R = (\sqrt{d} + 2\sqrt{r})^2$, we know by Theorem 1 that

$$\delta_\rho \;=\; 1 - \Pi(r,t,\rho) + 2e^{-cr}, \qquad \Pi(r,t,\rho) := \bar{\Phi}\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}}\right)\mathbb{P}(\chi_r^2 > t + \eta),$$

we further know that by this choice of R there exists a constant $C > 0$ s.t. $R = Cd$

$$\varepsilon_\rho \le \frac{d-r+1}{2}\log\Big(\rho + \sqrt{(1-\rho^2)\big(\tfrac{Cd^2}{t^2} - 1\big)}\Big) + \frac{C_1 d^3\sqrt{2(1-\rho)}}{2t^2},$$

where $C_1 = C^3$ is also a constant independent of d, r.

We first fix a target $\bar{\delta} \in (0,1)$ and we want to choose t in such a way that $\delta_\rho \leq \bar{\delta}$. This is equivalent to choosing t so that

$$\Pi(r,t,\rho) \geq 1 - \bar{\delta} + 2e^{-cr} = 1 - (\bar{\delta} - 2e^{-cr}) =: 1 - \bar{\delta}'$$

By construction $\bar{\rho}' < 1$, we further ensure that it does not become negative by clipping it at 0. Then by definition of $\Pi(r,t,\rho)$ we see that choosing t so that the following two conditions are fulfilled suffices:

$$\text{(i)} \quad \bar{\Phi}\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}}\right) \geq 1 - \bar{\delta}'/2$$
$$\text{(ii)} \quad \mathbb{P}(\chi_r^2 > t + \eta) \geq 1 - \bar{\delta}'/2$$

This is equivalent to

$$\text{(i)} \quad \bar{\delta}'/2 \geq \Phi\left(\frac{t(1-\rho) - \rho\eta}{\sqrt{(1-\rho^2)(t+\eta)}}\right)$$
$$\text{(ii)} \quad \bar{\delta}'/2 \geq \mathbb{P}(\chi_r^2 \leq t + \eta)$$

we define $z := \Phi^{-1}(\bar{\delta}'/2)$, note that for $\bar{\delta}' < 1$ we have $z < 0$ and $q := q_{\chi_r^2}(\frac{\bar{\delta}'}{2})$. So the first condition implies

$$t \leq \frac{(z\sqrt{1+\rho} + \sqrt{z^2(1+\rho) + 4\eta})^2}{4(1-\rho)} - \eta$$

and the second condition implies

$$t \leq q - \eta \quad \text{where} q \simeq r\left(1 - \frac{2}{9r} + z\sqrt{\frac{2}{9r}}\right)^3$$

Our goal is to choose η so that we can choose t as $O(r)$ while still fulfilling both bounds. There exists a sweet spot, as the first condition has a RHS increasing with η and the second has a RHS decreasing with η. And we need to fulfill $t \leq \min\{t_1(\eta), t_2(\eta)\}$. The best option is to choose η where the two meet:

$$t_1(\eta^\star) = t_2(\eta^\star)$$

Solving this allow us to choose $t_\star \simeq \rho r$ (up to $O(\sqrt{r})$ terms). Which give us that there exists a constant $\tilde{C} > 0$

$$\varepsilon_\rho \leq \tilde{C}\left(\frac{d-r+1}{2}\log\Big(\rho + \sqrt{(1-\rho^2)\big(\tfrac{d^2}{\rho^2 r^2} - 1\big)}\Big) + \frac{d^3\sqrt{2(1-\rho)}}{2\rho^2 r^2}\right).$$

□

Lemma 13 (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 $P_\perp = I - vv^\top$. Then,

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

We can write $z_i = vv^\top z_i + (I - vv^\top) z_i = v a_i + u_i$. Hence,

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

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

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

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

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

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

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

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

$$S \sim \sigma^2 \chi_r^2, \quad Y_\perp | (S = s) \overset{d}{=} N\left(0, \sigma^2 s U^\top P_\perp U\right) \overset{d}{=} N\left(0, \sigma^2 s I_{d-1}\right), \tag{10}$$

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

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

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

$$\mathbb{P}\left[S = s, Y_\perp = y_\perp\right] = \mathbb{P}\left[S = s\right] \mathbb{P}\left[Y_\perp = y_\perp | S = s\right] \tag{11}$$

As we know the distribution of S and $Y_\perp | S$ (Equation (10)), for $s > 0$,

$$\mathbb{P}\left[S = s\right] = f_{S\sigma^2}(s) = f_S\left(\frac{s}{\sigma^2}\right) = \frac{s^{r/2-1} e^{-\frac{s}{2\sigma^2}}}{2^{r/2} \Gamma(r/2) \sigma^{r-2}} \tag{12}$$

$$f_{Y_\perp | S}(y_\perp | S = s) = (2\pi)^{-\frac{d-1}{2}} \left(\sigma^2 s\right)^{-\frac{d-1}{2}} \exp\left(-\frac{\|y_\perp\|^2}{2\sigma^2 s}\right) \tag{13}$$

Substituting Equations (12) and (13) into Equation (11), we get

$$f_{S, Y_\perp}(s, y_\perp) = C_{r,d,\sigma} s^{\frac{r-d+1}{2}} e^{-\frac{s^2 + \|y_\perp\|^2}{2\sigma^2 s}}, \quad C_{r,d,\sigma} = \frac{1}{2^{r/2} \Gamma(r/2) \sigma^{d-r-1} (2\pi)^{(d-1)/2}} \tag{14}$$

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

$$Y = Q^\top \begin{pmatrix} S \\ Y_\perp \end{pmatrix}. \tag{15}$$

By changing the variables from $(S, Y_\perp)$ to Y with Equation (15), we get the probability density function (pdf) for Y when $y^\top v \geq 0$ ($s > 0$) from the pdf of $(S, Y_\perp)$ (Equation (14)), i.e.

$$\begin{aligned} f_Y(y) &= C_{r,d,\sigma} s^{\frac{r-d+1}{2}} \exp\left(-\frac{1}{2\sigma^2}\left(s + \frac{\left\|U^\top y\right\|^2}{v^\top y}\right)\right) \\ &= C_{r,d,\sigma}(v^\top y)^{\frac{r-d+1}{2}} \exp\left(-\frac{1}{2\sigma^2 v^\top y}\left(y^\top(vv^\top + UU^\top)y\right)\right) \\ &= C_{r,d,\sigma}(v^\top y)^{\frac{r-d+1}{2}} \exp\left(-\frac{\|y\|^2}{2\sigma^2 v^\top y}\right) \end{aligned} \tag{16}$$

This completes the proof. □

C Privacy of Projection Mechanism on matrices

Theorem 2. *For a dataset collection $\mathcal{D}$ and a query function f with outputs in $\mathbb{R}^{d\times m}$ of rank k, let ρ_k be the (column-space) alignment as in equation A1, and let ξ_k, ξ_1 be as in equation A2. Further, let $\Delta = \max_{S \overset{H}{\sim} S',\ S,S'\in\mathcal{D}} \|f(S) - f(S')\|_2$. Then for any tuning parameters $t > 0$, and $R > \xi_1(\sqrt{d} + \sqrt{r})^2$, the rank-r projection mechanism with $M = \sum_{i=1}^r z_i z_i^\top$ and $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d/r)$ is (ε, δ)-DP with*

$$\delta \leq e^{-\frac{1}{2}\left(\sqrt{k}-\sqrt{r}-(\sqrt{t}+\eta)/\xi_k\right)_+^2} + 2e^{-\frac{1}{2}\left(\frac{(\rho_k-1)t}{\sqrt{1-\rho_k^2}R} - \sqrt{m} - \sqrt{r}\right)_+^2} + 2e^{-c\left(\sqrt{\frac{R}{\xi_1}} - \sqrt{d} - \sqrt{r}\right)_+^2},$$

$$\varepsilon \leq \frac{(m-r)k}{2}\log\left(1 + \frac{2\xi_1}{\xi_k^2}\,\Delta\right) + \frac{r(d-r+1)}{2}\log\left(1 + \frac{R}{t}\,\Delta\right) + \frac{r^{3/2}\sqrt{k}\,R^3}{\sqrt{2}\,t^2}\Delta.$$

Proof. We first upper bound the likelihood ratio on a good set and then control the complement. Define the good set as

$$\mathcal{Y}_{t,R} \coloneqq \left\{Y \in \mathbb{R}^{d\times m} :\ \sigma_r\left(V^\top Y\right) > t,\ \sigma_r\left(V'^\top Y\right) > t,\ \|Y\|_2 \leq R\right\}. \tag{17}$$

Let $P(Y) \coloneqq f_{MV}(Y)$ and $Q(Y) \coloneqq f_{MV'}(Y)$. On $\mathcal{Y}_{t,R}$, $Q(Y) > 0$ ($\because \sigma_r\left(V'^\top Y\right) > t > 0$). For any measureable $\mathcal{Y} \subseteq \mathbb{R}^{d\times m}$, we define the maximum probability ratio on a good set as $L_{\max} = \max_{Y\in\mathcal{Y}_{t,R}} P(Y)/Q(Y)$. Then,

$$\frac{\mathbb{P}(MV \in \mathcal{Y} \cap \mathcal{Y}_{t,R})}{\mathbb{P}(MV' \in \mathcal{Y} \cap \mathcal{Y}_{t,R})} = \frac{\int_{Y\in\mathcal{Y}\cap\mathcal{Y}_{t,R}} P(Y)dY}{\int_{Y\in\mathcal{Y}\cap\mathcal{Y}_{t,R}} Q(Y)dY} \leq L_{\max}. \tag{18}$$

Similar to Equation (5),

$$\begin{aligned} \mathbb{P}(MV \in \mathcal{Y}) &= \mathbb{P}(MV \in \mathcal{Y} \cap \mathcal{Y}_{t,R}) + \mathbb{P}(MV \in \mathcal{Y} \cap \mathcal{Y}_{t,R}^c) \\ &\leq L_{\max}\mathbb{P}(MV' \in \mathcal{Y} \cap \mathcal{Y}_{t,R}) + \mathbb{P}(MV \in \mathcal{Y}_{t,R}^c) \\ &\leq L_{\max}\mathbb{P}(MV' \in \mathcal{Y}) + \mathbb{P}(MV \in \mathcal{Y}_{t,R}^c) \leq L_{\max}\mathbb{P}(MV' \in \mathcal{Y}) + \delta \end{aligned} \tag{19}$$

An analogous bound holds with V and V' swapped.

Therefore, we get ε by bounding the maximum likelihood ratio on the good set $L_{\max}$. By Lemma 14, we substitute in the pdf P, Q respectively, and get

$$L_{\max} \coloneqq \sup_{Y \in \mathcal{Y}_{t,R}} \frac{P(Y)}{Q(Y)}$$
$$= \sup_{Y \in \mathcal{Y}_{t,R}} \left(\underbrace{\frac{\left|V^\top V\right|_+}{\left|V'^\top V'\right|_+}}_{\text{Part I}} \right)^{\frac{m-r}{2}} \left(\underbrace{\frac{\left|V'^\top Y\right|_+}{\left|V^\top Y\right|_+}}_{\text{Part II}} \right)^{\frac{d+1-r}{2}} \exp\left(\frac{r}{2} \underbrace{\operatorname{Tr}\left(\left[\left(V'^\top Y\right)^\dagger - \left(V^\top Y\right)^\dagger \right] Y^\top Y \right)}_{\text{Part III}} \right).$$

Part I. Under equation A2, $\sigma_k(V), \sigma_k(V') \geq \xi_k > 0$ and $\sigma_1(V), \sigma_1(V') \leq \xi_1 < \infty$. Then

$$\begin{aligned} \log \frac{\left|V^\top V\right|_+}{\left|V'^\top V'\right|_+} &= \sum_{i=1}^{k} \log \frac{\lambda_i(V^\top V)}{\lambda_i(V'^\top V')} = \sum_{i=1}^{k} \log\left(1 + \frac{\lambda_i(V^\top V) - \lambda_i(V'^\top V')}{\lambda_i(V'^\top V')}\right) \\ &\leq k \log\left(1 + \frac{\left\|V^\top V - V'^\top V'\right\|_2}{\xi_k^2}\right) \\ &\leq k \log\left(1 + \frac{2\xi_1}{\xi_k^2} \left\|V - V'\right\|_2\right), \end{aligned} \tag{20}$$

where the first inequality follows from Weyl's inequality (Lemma 11) and the second from $\|V^\top V - V'^\top V'\|_2 \leq \|V^\top (V - V')\|_2 + \|(V - V')^\top V'\|_2 \leq 2\xi_1 \|V - V'\|_2$.

Part II. On $\mathcal{Y}_{t,R}$, $\sigma_r(V'^\top Y) \geq t$ and $\|Y\|_2 \leq R$. By definition, M is rank r. Then, $Y = MV$ is at most rank r, and both $V^\top Y$ and $V'^\top Y$ are at most rank r. Hence,

$$\begin{aligned} \log \frac{\left|V'^\top Y\right|_+}{\left|V^\top Y\right|_+} &= \sum_{i=1}^{r} \log \frac{\sigma_i(V'^\top Y)}{\sigma_i(V^\top Y)} = \sum_{i=1}^{r} \log\left(1 + \frac{\sigma_i(V'^\top Y) - \sigma_i(V^\top Y)}{\sigma_i(V^\top Y)}\right) \\ &\leq r \log\left(1 + \frac{\left\|V'^\top Y - V^\top Y\right\|_2}{t}\right) \\ &\leq r \log\left(1 + \frac{\|V' - V\|_2 \|Y\|_2}{t}\right) \leq r \log\left(1 + \frac{R}{t} \left\|V' - V\right\|_2\right). \end{aligned} \tag{21}$$

where the first inequality follows from Lemma 11 and the definition of good set (Equation (17)).

Part III. Using Cauchy–Schwarz and a standard pseudoinverse perturbation bound (Lemma 10),

$$\begin{aligned} \operatorname{Tr}\left(\left(\left(V'^\top Y\right)^\dagger - \left(V^\top Y\right)^\dagger \right) Y^\top Y \right) &\leq \left\| \left(V'^\top Y\right)^\dagger - \left(V^\top Y\right)^\dagger \right\|_F \|Y\|_F^2 \\ &\leq \frac{\sqrt{2} \left\| (V' - V)^\top Y \right\|_F}{\sigma_r(V^\top Y) \sigma_r\left(V'^\top Y\right)} \|Y\|_F^2 \\ &\leq \frac{\sqrt{2} \|V' - V\|_2 \|Y\|_F^3}{t^2} \\ &\leq \frac{\sqrt{2} r^{3/2} R^3}{t^2} \left\|V' - V\right\|_2, \end{aligned} \tag{22}$$

since $\|Y\|_F \leq \sqrt{r} \|Y\|_2 \leq \sqrt{r} R$ on $\mathcal{Y}_{t,R}$.

Combining Parts I–III yields the stated upper bound on ε by substituting $\Delta \coloneqq \|V' - V\|_2$.

It remains to bound the bad set probability $\mathbb{P}\left(\mathcal{Y}_{t,R}^c\right)$. Let $S \coloneqq V^\top M V$ and $S'' \coloneqq V'^\top M V'$ (both are symmetric), and define the events

$$\mathcal{E}_1 \coloneqq \{\lambda_r(S) \le t\}, \quad \mathcal{E}_2 \coloneqq \left\{\lambda_r(S'') \le t\right\}, \quad \mathcal{E}_3 \coloneqq \{\|MV\|_2 > R\}, \quad \tilde{\mathcal{E}}_3 \coloneqq \{\|M\|_2 > R/\xi_1\}.$$

Then $\mathcal{E}_3 \subseteq \tilde{\mathcal{E}}_3$ since $\|MV\|_2 \le \|M\|_2 \|V\|_2 \le \xi_1 \|M\|_2$. Hence

$$\mathbb{P}\left(\mathcal{Y}_{t,R}^c\right) = \mathbb{P}\left(\mathcal{E}_1 \cup \mathcal{E}_2 \cup \mathcal{E}_3\right) \le \mathbb{P}(\mathcal{E}_1) + \mathbb{P}(\mathcal{E}_2) + \mathbb{P}(\tilde{\mathcal{E}}_3).$$

Bounding $\mathcal{E}_1$. Write $M = \frac{1}{r}\widetilde{H}\widetilde{H}^\top$ with $\widetilde{H} \in \mathbb{R}^{d\times r}$ i.i.d. standard Gaussian. Let $V = U\Lambda W^\top$ be an SVD with $\Lambda = \operatorname{diag}(\sigma_1, \dots, \sigma_k, 0, \dots)$. Then

$$S = V^\top M V = \frac{1}{r} W \Lambda^\top (U^\top \widetilde{H})(U^\top \widetilde{H})^\top \Lambda W^\top,$$

so by Lemma 5, $\lambda_r(S) = \frac{1}{r}\sigma_r(\Lambda^\top U^\top \widetilde{H})^2$. With $G \coloneqq U^\top \widetilde{H}$ and the projector P onto the first k coordinates,

$$\sigma_r(\Lambda^\top G) \ \ge \xi_k \sigma_r(PG) = \xi_k \sigma_r(\tilde{G}),$$

where $\tilde{G} \in \mathbb{R}^{k\times r}$ is i.i.d. standard Gaussian. Hence, using Lemma 8,

$$\mathbb{P}\left(\lambda_r(S) \le t\right) = \mathbb{P}\left(\sigma_r(\tilde{G}) \le \sqrt{rt}/\xi_k\right) \le \exp\left(-\tfrac{1}{2}\left(\sqrt{k} - \sqrt{r} - t/\xi_k\right)_+^2\right). \tag{23}$$

Bounding $\mathcal{E}_2$. Write $M = \frac{1}{r}\widetilde{H}\widetilde{H}^\top$, set $Y \coloneqq V^\top \widetilde{H}$ and $Z \coloneqq V'^\top \widetilde{H}$. Then

$$\mathbb{P}\left\{\lambda_r(V^\top M V) \ge t,\ \lambda_r(V'^\top M V) \ge t\right\} = \mathbb{E}\left[\mathbf{1}_{\sigma_r(Y)>\sqrt{t}}\, \mathbb{E}\left[\mathbf{1}_{\sigma_r(ZY)>\sqrt{t}} \mid Y\right]\right].$$

To bound the inner term, we first denote $P \coloneqq V(V^\top V)^\dagger V^\top$ and $\Theta \coloneqq (V^\top V)^\dagger V^\top V'$, and use the conditional representation

$$ZY^\top \mid Y = \Theta^\top Y Y^\top + S^{1/2} G Y^\top,$$

where $S \coloneqq V'^\top (I - P) V'$ and $G \in \mathbb{R}^{m\times r}$ is distributed i.i.d. normal. So by the triangle inequality for singular values,

$$\sigma_r(ZY^\top) \ \ge\ \sigma_{\min}(\Theta^\top Y)\sigma_r(Y) - \left\|S^{1/2}\right\|_2 \|G\|_2 \|Y\| \ge \sigma_{\min}(\Theta|_{\operatorname{span}(Y)})\sigma_r^2(Y) - \left\|S^{1/2}\right\| \|G\| \|Y\|.$$

where

$$\sigma_{\min}\left(\Theta|_{\operatorname{span}(Y)}\right) = \inf_{u\in\operatorname{span}(Y):\|u\|=1} \|\Theta u\| \ge \rho_k.$$

Hence, on the event $\left\{Y : \sigma_r(Y) \ge \sqrt{t} + \eta\right\}$,

$$\begin{aligned}\mathbb{P}\left(\sigma_r(ZY^\top) \le t \mid Y\right) &\le \mathbb{P}\left(\|G\|_2 \ge \frac{\sigma_{\min}(\Theta|_{\operatorname{span}(Y)})(\sqrt{t}+\eta)^2 - t}{\left\|S^{1/2}\right\| \|Y\|}\right) \\ &\le 2\exp\left(-\tfrac{1}{2}\left(\frac{\rho_k(\sqrt{t}+\eta)^2 - t}{\sqrt{1-\rho_k^2}\|Y\|} - \sqrt{m} - \sqrt{r}\right)_+^2\right),\end{aligned}$$

using $\sigma_{\min}(\Theta|_{\mathrm{span}(Y)}) \geq \rho_k$,

$$\|S^{1/2}\|_2 = \sin\theta_{\max} = \sqrt{1-\cos^2\theta_{\max}} = \sqrt{1-\sigma_k(\Theta)^2} = \sqrt{1-\rho_k^2},$$

and the Gaussian spectral–norm tail (Lemma 7) $\mathbb{P}\left(\|G\|_2 \geq \sqrt{m}+\sqrt{r}+u\right) \leq 2e^{-u^2/2}$.

Substituting the conditional bound into the expansion gives

$$\begin{aligned}
&\mathbb{E}\left[\mathbf{1}_{\sigma_r(Y)>\sqrt{t}}\mathbb{E}\left\{\mathbf{1}_{\sigma_r(Z)>\sqrt{t}} \mid Y\right\}\right] \\
&\quad \geq \mathbb{P}\left(\sigma_r(Y) > \sqrt{t}+\eta\right)\left(1-2\exp\left(-\tfrac{1}{2}\left(\tfrac{\rho_k(\sqrt{t}+\eta)^2-t}{\sqrt{1-\rho_k^2}R}-\sqrt{m}-\sqrt{r}\right)_+^2\right)\right),
\end{aligned}$$

because the inner probability is non-decreasing in $\sigma_r(Y)$ and we restrict to $\{\sigma_r(Y) \geq \sqrt{t}+\eta\}$. Now use the standard smallest–singular–value tail (Lemma 8) for $Y = V^\top \widetilde{H}$ with $\sigma_k(V) \geq \xi_k$:

$$\mathbb{P}\left(\sigma_r(Y) \leq \sqrt{t}+\eta\right) \leq \exp\left(-\tfrac{1}{2}\left(\sqrt{k}-\sqrt{r}-(\sqrt{t}+\eta)/\xi_k\right)_+^2\right),$$

so

$$\begin{aligned}
&\mathbb{P}\left\{\lambda_r(V^\top M V) \geq t,\ \lambda_r(V'^\top M V) \geq t\right\} \\
&\geq \left(1-e^{-\frac{1}{2}\left(\sqrt{k}-\sqrt{r}-(\sqrt{t}+\eta)/\xi_k\right)_+^2}\right)\left(1-2e^{-\frac{1}{2}\left(\frac{\rho_k(\sqrt{t}+\eta)^2-t}{\sqrt{1-\rho_k^2}R}-\sqrt{m}-\sqrt{r}\right)_+^2}\right) \\
&\geq 1-e^{-\frac{1}{2}\left(\sqrt{k}-\sqrt{r}-(\sqrt{t}+\eta)/\xi_k\right)_+^2}-2e^{-\frac{1}{2}\left(\frac{\rho_k(\sqrt{t}+\eta)^2-t}{\sqrt{1-\rho_k^2}R}-\sqrt{m}-\sqrt{r}\right)_+^2}.
\end{aligned}$$

Bounding $\tilde{\mathcal{E}}_3$. Since $M = \frac{1}{r}\widetilde{H}\widetilde{H}^\top$, we have $\|M\|_2 = \frac{1}{r}\left\|\widetilde{H}\right\|_2^2$. Using Lemma 7, we obtain

$$\mathbb{P}\left(\tilde{\mathcal{E}}_3\right) = \mathbb{P}\left(\|M\|_2 > R/\xi_1\right) \leq 2\exp\left(-c\left(\sqrt{\tfrac{R}{\xi_1}}-\sqrt{d}-\sqrt{r}\right)_+^2\right). \tag{24}$$

Putting the three bounds together yields the stated δ.

$$\begin{aligned}
\mathbb{P}(\tilde{\mathcal{E}}_1^c \cup \tilde{\mathcal{E}}_2^c \cup \tilde{\mathcal{E}}_3^c) &\leq \mathbb{P}(\tilde{\mathcal{E}}_1^c \cup \tilde{\mathcal{E}}_2^c) + \mathbb{P}(\tilde{\mathcal{E}}_3^c) = 1-\mathbb{P}(\tilde{\mathcal{E}}_1 \cup \tilde{\mathcal{E}}_2) + \mathbb{P}(\tilde{\mathcal{E}}_3^c) \\
&\leq \exp\left(-\frac{1}{2}\left(\sqrt{k}-\sqrt{r}-\left(\sqrt{t}+\eta\right)/\xi_k\right)_+^2\right) \\
&\quad + 2\exp\left(-\frac{1}{2}\left(\frac{\rho_k(\sqrt{t}+\eta)^2-t}{\sqrt{1-\rho_k^2}R}-\sqrt{m}-\sqrt{r}\right)_+^2\right) \\
&\quad + 2\exp\left(-c\left(\frac{\sqrt{R}}{\xi_1}-\sqrt{d}-\sqrt{r}\right)_+^2\right)
\end{aligned}$$

□

Lemma 14 (PDF of MV when $V \in \mathbb{R}^{d\times m}$)**.** *Let $z_1, ..., z_r$ be i.i.d. $\mathcal{N}(0, \mathbf{I}_d/r)$ and $M = \sum_{i=1}^r z_i z_i^T$. For $V \in \mathbb{R}^{d\times m}$ of rank k with $r \le k \le \min(m, d)$, the density of $Y = MV$ is*

$$f_{MV}(Y) = \begin{cases} C_{m,k,r,d} \left|V^\top V\right|_+^{\frac{m-r}{2}} |S|_+^{\frac{r-d-1}{2}} \exp\left(-\frac{r \operatorname{Tr}\left(S^\dagger Y^\top Y\right)}{2}\right), & \text{if } S = V^\top Y \succeq 0 \text{ \& } rank(S) = r, \\ 0, & \text{otherwise} \end{cases}$$

for $C_{m,k,r,d} = \frac{r^{\frac{m(r+d-m)}{2}}}{(2\pi)^{m(d-k)} 2^{\frac{mr}{2}} \Gamma_m\left(\frac{r}{2}\right)}$

Remark 1. *For a symmetric matrix A with eigenvalues $\{\lambda_i\}$ define*

$$|A|_+ := \prod_{\lambda_i > 0} \lambda_i$$

Proof of Lemma 14. Let $S = V^\top MV$. Then $S \sim \frac{1}{r}\text{Wishart}_m(r, V^\top V)$. Write $z_i = Ba_i + u_i$ with $a_i := V^\top z_i,\ u_i := P_\perp z_i,\ B := V\left(V^\top V\right)^\dagger,\ P_\perp := I - V(V^\top V)^\dagger V^\top$. Then

$$Y = \sum_{i=1}^r z_i a_i^\top = BS + UA,$$

where $U = [u_1, \ldots, u_r] \in \mathbb{R}^{d\times r}$ and $A = [a_1^\top; \ldots; a_r^\top] \in \mathbb{R}^{r\times m}$ with $S = A^\top A$. One can check $U \perp A$, and conditionally on S, $UA \sim \text{MN}_{d\times m}(0, \frac{1}{r}P_\perp, S)$. Here, MN denotes the *matrix normal distribution*.

Let $\Pi_\perp \in \mathbb{R}^{d\times(d-k)}$ be an orthonormal basis of $\text{col}(P_\perp)$ and set $Y_\perp := \Pi_\perp^\top Y$. Then $Y_\perp \mid S \sim \text{MN}_{(d-k)\times m}(0, \frac{1}{r}I_{d-k}, S)$. The change of variables $T(Y) := (S, Y_\perp) = (V^\top Y, \Pi_\perp^\top Y)$ has Jacobian determinant $\left|V^\top V\right|_+^{m/2}$. Using the Wishart density for S, the matrix-normal density for $Y_\perp \mid S$, and change of variables formula we obtain

$$f_Y(y) = C_{m,k,r,d} \left|V^\top V\right|_+^{\frac{m-r}{2}} |S|_+^{\frac{r-d-1}{2}} \exp\left(-\frac{r}{2}\left(\operatorname{Tr}\left(\left(V^\top V\right)^\dagger S\right) + \operatorname{Tr}\left(S^\dagger Y_\perp^\top Y_\perp\right)\right)\right).$$

Since $Y_\perp^\top Y_\perp = Y^\top Y - Y^\top V(V^\top V)^\dagger V^\top Y = Y^\top Y - S^\top (V^\top V)^\dagger S$ and $S = S^\top$, we have

$$\operatorname{Tr}\left(\left(V^\top V\right)^\dagger S\right) - \operatorname{Tr}\left(S^\dagger S\left(V^\top V\right)^\dagger S\right) = \operatorname{Tr}\left(\left(V^\top V\right)^\dagger \left(SS^\dagger S - S\right)\right) = 0,$$

using $SS^\dagger S = S$. Therefore the exponent simplifies to $-\frac{r}{2}\operatorname{Tr}(S^\dagger Y^\top Y)$, as claimed. □

D Privacy amplification and applications

Lemma 4. *Let $v, v' \in \mathbb{R}^d$ be two unit vectors with with $\cos\angle(v, v') = v^\top v' \ge \rho$, $z \in \mathcal{N}(0, \mathbf{I}_d)$, $\delta > 0$ and $\gamma > \frac{1-\rho}{1+\rho}\sqrt{\frac{2}{d}\log\frac{8}{\delta}}$, then with probability at least $1 - \delta$, we have*

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

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

Proof. Given two firxed unit norm vectors v and v' with alignment $\rho := v^\top v'$, we want to show that $\tilde{v} := v + \frac{\gamma z}{\|z\|_2}$ and $\tilde{v}' := v' + \frac{\gamma z}{\|z\|_2}$ are more aligned than v and v' with high probability. Note we assume $-1 \le \rho < 1$ as $\rho = 1$ means $v = v'$ and therefore is a trivial case.

Let $u = \frac{z}{\|z\|_2}$ be a uniform vector from the d-dimensional unit Euclidean ball. Then, we can rewrite $\tilde{v} = v + \gamma u$ and $\tilde{v}' = v' + \gamma u$ and

$$\begin{aligned}
\cos(\tilde{v}, \tilde{v}') &= \cos\left(v + \frac{\gamma z}{\|z\|_2} + \gamma u\right) \\
&= \frac{\langle v + \gamma u, v' + \gamma u\rangle}{\|v + \gamma u\| \, \|v' + \gamma u\|} \\
&\ge \frac{\gamma^2 + \rho + \gamma\left(\langle v, u\rangle + \langle v', u\rangle\right)}{\sqrt{1 + \gamma^2 + \gamma\langle v, u\rangle}\sqrt{1 + \gamma^2 + \gamma\langle v', u\rangle}}.
\end{aligned}$$

By Lemma 9, setting $t = \sqrt{\frac{2}{d}\ln\frac{8}{\delta}}$

$$\begin{aligned}
\mathbb{P}\left[\langle v, u\rangle \ge \sqrt{\frac{2}{d}\ln\frac{8}{\delta}}\right] \le \frac{\delta}{4}, \quad \mathbb{P}\left[\langle v', u\rangle \ge \sqrt{\frac{2}{d}\ln\frac{8}{\delta}}\right] \le \frac{\delta}{4} \\
\mathbb{P}\left[\langle v, u\rangle \le -\sqrt{\frac{2}{d}\ln\frac{8}{\delta}}\right] \le \frac{\delta}{4}, \quad \mathbb{P}\left[\langle v, u\rangle \le -\sqrt{\frac{2}{d}\ln\frac{8}{\delta}}\right] \le \frac{\delta}{4}
\end{aligned}$$

Therefore, with probability at least $1 - \delta$,

$$\sqrt{\frac{2}{d}\ln\frac{8}{\delta}} \le \langle v, u\rangle \le \sqrt{\frac{2}{d}\ln\frac{8}{\delta}}, \quad -\sqrt{\frac{2}{d}\ln\frac{8}{\delta}} \le \langle v', u\rangle \le \sqrt{\frac{2}{d}\ln\frac{8}{\delta}} \tag{25}$$

Which means with probability at least $1 - \delta$

$$\cos\left(v + \gamma u, v' + \gamma u\right) \ge \frac{\rho - \gamma s + \gamma^2}{1 + \gamma s + \gamma^2}$$

where $s = 2\sqrt{\frac{2}{d}\ln\frac{8}{\delta}}$. Subtracting ρ from both sides and simplifies the equation, we complete the proof. □

E Additional results on convergence guarantee of RP-GD

In this section, we analyse the convergence guarantee of projection mechanism when applied to convex optimization problem. We show that under the minimum alignment assumption between the dataset collection $\mathcal{D}$ and the gradient trajectory f, gradient descent with projection mechanism (RP-GD) achieves a convergence rate comparable to, and in some regimes better than, that of DP-GD for smooth convex optimization.

For a dataset $S = \{(x_i, y_i)\}_{i=1}^n$, let $\ell : \mathcal{W} \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$ be a loss function. Then, let $\mathcal{L} : \mathcal{W} \times \mathcal{Z} \to \mathbb{R}_+$ be the loss function over the entire dataset, i.e. $\mathcal{L}(w; S) = \frac{1}{n}\sum_{i=1}^n \ell(w, (x_i, y_i))$. The optimization problem is defined in Equation (26).

$$\hat{w} = \arg\min_{w \in \mathcal{W}} \frac{1}{n}\sum_{i=1}^n \ell(w, x_i, y_i) =: \arg\min_{w \in \mathcal{W}} \mathcal{L}(w; S). \tag{26}$$

The most common private approach is DP-GD (equivalently, Noisy-GD when there is no gradient clipping), which attains a convergence rate of $\sqrt{d}/(T\varepsilon)$ [Abadi et al., 2016, Bassily et al., 2014]. At each step, DP-GD applies the Gaussian mechanism (Lemma 1) to the gradient and then performs a gradient-descent update with the noisy gradient. In contrast, we propose *Randomly Projected Gradient Descent (RP-GD)*: first apply the projection mechanism (Definition 1) to the gradient, then take a gradient descent step with the projected output. Specifically, sample $M \sim \mathsf{W}_d(\sigma^2 I_d, r)$ once and update at each time step with the following update function,

$$w_{t+1} = w_t - \eta M \nabla \mathcal{L}(w_t). \tag{27}$$

RP-GD enjoys the following convergence guarantee.

Theorem 3. *For a dataset collection $\mathcal{D}$ and the query function $f(w_0, S) = \sum_{i=0}^{T} \nabla \mathcal{L}(w_i; S)$, let the minimum alignment of f on $\mathcal{D}$ be ρ. For any dataset $S \in \mathcal{D}$, if $d \leq \frac{3}{4}r$, $r \geq 16 \log \frac{1}{\delta}$ and assume $\mathcal{L}$ is convex and β-smooth, then RP-GD is $(\varepsilon_\rho, \delta)$-DP where $\varepsilon_\rho = Cr\sqrt{1-\rho}/\rho$ for some constant C independent of d, r, and with probability at least $1 - 2\delta$,*

$$\mathcal{L}(w_T) - \mathcal{L}_S^\star = O\left(\frac{\beta \left\|w_0 - w_S^\star\right\|_2}{2T}\right)$$

where $w_S^\star = \arg\min \mathcal{L}(w; S)$ and $\mathcal{L}_S^\star = \min \mathcal{L}(w; S)$.

Proof. **Convergence:** For a dataset $S \in \mathcal{D}$, condition on a projection matrix M with $\lambda_{max}(M) > 0$ and $\lambda_{\min}(M) > 0$. We first show that $\mathcal{L}(w_T) - \mathcal{L}_S^\star \leq \frac{\beta \lambda_{\max} \left\|w_0 - w_S^\star\right\|_2}{T \lambda_{\min}}$.

Let $v_t = \nabla \mathcal{L}(w_t)$. By smoothness of $\mathcal{L}$,

$$\begin{aligned}
\mathcal{L}(w_{t+1}) &\leq \mathcal{L}(w_t) - \eta v_t^\top M v_t + \frac{\beta \eta^2}{2} \left\|M v_t\right\|^2 \\
&\leq \mathcal{L}(w_t) - \eta v_t^\top M v_t + \frac{\beta \eta^2}{2} \lambda_{\max} v_t^\top M v_t \\
&\leq \mathcal{L}(w_t) - \eta \left(1 - \frac{\beta \eta \lambda_{\max}}{2}\right) v_t^\top M v_t \\
&\leq \mathcal{L}(w_t) - \frac{\eta}{2} v_t^\top M v_t
\end{aligned} \tag{28}$$

where the last inequality follows by $\eta \leq \frac{1}{2\beta\lambda_{\max}}$. Rearranging,

$$v_t^\top M v_t \leq \frac{2}{\eta} \left(\mathcal{L}(w_t) - \mathcal{L}(w_{t+1})\right) \tag{29}$$

Then

$$\begin{aligned}
\left\|w_{t+1} - w_S^\star\right\|_{M^\dagger}^2 &= (w_{t+1} - w_S^\star)^\top M^\dagger (w_{t+1} - w_S^\star) \\
&= (w_t - w_S^\star - \eta M v_t)^\top M^\dagger (w_t - w_S^\star - \eta M v_t) \\
&= (w_t - w_S^\star)^\top M^\dagger (w_t - w_S^\star) - 2\eta v_t^\top M M^\dagger (w_t - w_S^\star) + \eta^2 v_t^\top M v_t \\
&= \left\|w_t - w_S^\star\right\|_{M^\dagger} - 2\eta v_t^\top (w_t - w_S^\star) - 2\eta v_t^\top (M M^\dagger - I)(w_t - w_S^\star) + \eta^2 v_t^\top M v_t.
\end{aligned} \tag{30}$$

By convexity of $\mathcal{L}$,

$$\mathcal{L}(w_t) - \mathcal{L}(w_S^\star) \leq v_t^\top (w_t - w_S^\star)$$

Therefore, by convexity and Equation (29)

$$\begin{aligned}\|w_{t+1} - w_S^\star\|_{M^{-1}}^2 &\le \|w_t - w_S^\star\|_{M^{-1}} - 2\eta\left(\mathcal{L}(w_t) - \mathcal{L}(w_S^\star)\right) + 2\eta\left(\mathcal{L}(w_t) - \mathcal{L}(w_{t+1})\right) - 2\eta v_t^\top (MM^\dagger - I)(w_t - w_S^\star) \\ &= \|w_t - w_S^\star\|_{M^{-1}} + 2\eta\left(\mathcal{L}(w_S^\star) - \mathcal{L}(w_{t+1})\right) + 2\eta v_t^\top (I - MM^\dagger)(w_t - w_S^\star)\end{aligned} \tag{31}$$

Rearrange,

$$\mathcal{L}(w_{t+1}) - \mathcal{L}(w_S^\star) \le \frac{1}{2\eta}\left(\|w_t - w_S^\star\|_{M^{-1}} - \|w_{t+1} - w_S^\star\|_{M^{-1}}^2\right) + v_t^\top (M^\dagger M - I)(w_0 - w_S^\star) \tag{32}$$

where the last part follows by $w_t = w_0 + M\sum_{i=1}^t \eta v_t$ and $(M^\dagger M - I)M = 0$.

As $\mathcal{L}(w_t)$ is monotonically decreasing (Equation (29)),

$$\begin{aligned}\mathcal{L}(w_T) - \mathcal{L}(w_S^\star) &\le \frac{1}{T}\sum_{t=1}^T \mathcal{L}(w_t) - \mathcal{L}(w_S^\star) \\ &\le \frac{1}{2\eta T}\sum_{t=1}^T \|w_t - w_S^\star\|_{M^{-1}} - \|w_{t+1} - w_S^\star\|_{M^\dagger}^2 + \frac{1}{T}\sum_{i=1}^T v_t^\top (M^\dagger M - I)(w_0 - w_S^\star) \quad \because \textit{Equation (32)} \\ &\le \frac{\|w_0 - w_S^\star\|_{M^\dagger}}{2\eta T} \le \frac{\|w_0 - w_S^\star\|\,\lambda_{\max}(M)}{2T\lambda_{\min}(M)}\end{aligned} \tag{33}$$

By initializing $w_0 \sim \mathcal{N}(0, I_d/(dT))$, the first term $\left\|(M^\dagger M - I)w_0\right\| = O\left(\frac{1}{Td}\right)$ with high probability. (Variance σ^2 cancels out) By Lemma 8 and Lemma 7, for some t, t', and for $M = \sigma^2 ZZ^\top$ where $Z \in \mathbb{R}^{d\times r}$ with each entry i.i.d. standard Gaussian

$$\mathbb{P}(\lambda_{\max}(M) > \sigma^2(\sqrt{d} + \sqrt{r} + t)^2) = \mathbb{P}(\lambda_{\max}(Z) > \sqrt{d} + \sqrt{r} + t) \le e^{-\frac{t^2}{2}}$$

Setting $t = \sqrt{2\log\frac{1}{\varsigma}}$, we have

$$\mathbb{P}\left[\lambda_{\max}(M) > \left(\sqrt{d} + \sqrt{r} + \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] \le \varsigma$$

As $r \ge 3d/4$,

$$\begin{aligned}\mathbb{P}\left[\lambda_{\max}(M) > \sigma^2\left(2\sqrt{r} + \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] &\le \mathbb{P}\left[\lambda_{\max}(M) > \sigma^2\left(\left(1 + \frac{\sqrt{3}}{2}\right)\sqrt{r} + \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] \\ &\le \mathbb{P}\left[\lambda_{\max}(M) > \sigma^2\left(\sqrt{d} + \sqrt{r} + \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] \le \varsigma\end{aligned} \tag{34}$$

Similarly,

$$\mathbb{P}\left[\lambda_{\min}(M) < \sigma^2\left(\frac{1}{2}\sqrt{r} - \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] \le \mathbb{P}\left[\lambda_{\min}(M) < \sigma^2\left(\sqrt{r} - \sqrt{d} - \sqrt{2\log\frac{1}{\varsigma}}\right)^2\right] \le \varsigma \tag{35}$$

Condition on these two events, which occur with probability at least $1 - 2\varsigma$, we substitute in $\lambda_{\max} \le \sigma^2\left(2\sqrt{r} + \sqrt{2\log\frac{1}{\varsigma}}\right)^2$ and $\lambda_{\min} \ge \sigma^2\left(\frac{1}{2}\sqrt{r} - \sqrt{2\log\frac{1}{\varsigma}}\right)^2$, which yields the desired result. □

Our bound improves upon DP-GD when the dataset collection has well-aligned gradient sums. In particular, when the minimum alignment is large (e.g., $\rho = \frac{r}{r+1}$), the privacy parameter ε remains a constant, and RP-GD achieves an $O(1/T)$ convergence rate that is independent of the dimension d. By contrast, if we fix $\varepsilon = C$, then DP-GD converges at rate $O(\sqrt{d}/T)$. For less aligned datasets, for example, when $0 < \rho < 0.5$, we obtain $\varepsilon = O(d)$, and the convergence guarantee of RP-GD is comparable to that of DP-SGD.

If the minimum-alignment assumption does not hold for the cumulative gradient over T steps, we can relax it by requiring alignment only at the per-step gradient level and resampling a new projection at each step. This relaxation incurs an additional factor of T in the privacy guarantee due to composition across steps. Notably, when $r \leq d$, this procedure coincides with CompSGD [Kasiviswanathan, 2021], which also achieves nearly dimension-independent convergence guarantees.

Low-rank version also saves space and computation We note that Equation (27) is equivalent to first performing a gradient descent in random projected subspaces using a variable $z_0 = Zw_0 \in \mathbb{R}^r$ and the projection matrix $Z \in \mathbb{R}^{d\times r}$ where $M = ZZ^\top$, by parameterizing $w_t = w_t + Z^\top z_t$ [Hao et al., 2024]. Therefore, when $r < d$, RP-GD is also more computationally and memory efficient.